

# Optimal graphon estimation in cut distance

Olga Klopp\* and Nicolas Verzelen†

March 16, 2017

## Abstract

Consider the twin problems of estimating the connection probability matrix of an inhomogeneous random graph and the graphon of a  $W$ -random graph. We establish the minimax estimation rates with respect to the cut metric for classes of block constant matrices and step function graphons. Surprisingly, our results imply that, from the minimax point of view, the raw data, that is, the adjacency matrix of the observed graph, is already optimal and more involved procedures cannot improve the convergence rates for this metric. This phenomenon contrasts with optimal rates of convergence with respect to other classical distances for graphons such as the  $l_1$  or  $l_2$  metrics.

*Keywords:* inhomogeneous random graph, graphon,  $W$ -random graphs, networks, stochastic block model, cut distance.

## 1 Introduction

In the last decade, network analysis has become an important research field driven by applications in social sciences, computer sciences, statistical physics, genomics, ecology... A flourishing line of literature amounts to fit observed networks to parametric or non-parametric models of random graphs. Among the parametric models, one of the most popular is the stochastic block model [23]. In the stochastic block model with  $n$  vertices and  $k$  blocks, the class  $Z_i$  of each vertex  $i \in [n]$  is drawn independently in  $[k]$  according to some probability distribution  $\pi$ . Given  $Z$ , the edges of the graph are then sampled independently, the probability that there is an edge between  $i$  and  $j$  being equal to  $Q_{Z_i Z_j}$  where  $Q = (Q_{ij}) \in [0, 1]^{k \times k}$  is a given symmetric matrix. Although this model is suitable for analyzing small networks, it does not allow to analyze the finer structures of extremely large networks. To go beyond the possible limitation of parametric models, non-parametric models of random graphs have been introduced [18, 22].

One possible non-parametric generalization of the stochastic block models is given by the  $W$ -random graph model [18] based on the notion of graphon. Graphons are symmetric measurable functions  $W : [0, 1]^2 \rightarrow [0, 1]$ . In the sequel, the space of graphons is denoted by  $\mathcal{W}^+$ . Given a graphon  $W_0 \in \mathcal{W}^+$ , a graph on  $n$  vertices is sampled according to the  $W$ -random graph model in the following way. Let  $\Theta_0 = (\Theta_{ij})$  be a  $n \times n$  random symmetric matrix defined by

$$\Theta_{ij} = \rho_n W_0(\xi_i, \xi_j), \forall i \neq j \text{ and } \Theta_{ii} = 0 \quad (1)$$

where  $1 \geq \rho_n > 0$  is the scale parameter that can be interpreted as the expected proportion of non-zero edges and  $\xi_1, \dots, \xi_n$  are unobserved (latent) i.i.d. random variables uniformly distributed

\*CREST and MODAL'X, University Paris Nanterre, FRANCE, [olga.klopp@math.cnrs.fr](mailto:olga.klopp@math.cnrs.fr)

†INRA, UMR 729 MISTEA, F-34060 Montpellier, FRANCE, [nicolas.verzelen@inra.fr](mailto:nicolas.verzelen@inra.fr)

on  $[0, 1]$ . Then, given  $\Theta_0$ , the graph is sampled according to the inhomogeneous random graph model [6]. More precisely, vertices  $i$  and  $j$  are connected by an edge with probability  $\Theta_{ij}$  and these events are independent for all pairs  $(i, j)$  with  $i < j$ . When  $\Theta_0$  is considered as a deterministic matrix, we call it inhomogeneous random graph model with respect to  $\Theta_0$ . If  $W_0$  is a step-function with  $k$  steps, the graph is distributed as a stochastic block model with  $k$  groups. The case of a dense graph corresponds to  $\rho_n = 1$ , whereas the choice  $\rho_n \rightarrow 0$  when  $n \rightarrow \infty$  produces sparser graphs. This model was recently studied by a number of authors, see e.g., [4, 5, 16, 17, 27, 34].

In the present paper we consider the problems of estimating the matrix of connection probabilities  $\Theta_0$  and the graphon  $f_0 = \rho_n W_0$  from a single observation of a graph. Suppose that we observe the  $n \times n$  adjacency matrix  $\mathbf{A} = (A_{ij})$  of a graph that has either been sampled according to the inhomogeneous random graph model with a fixed matrix  $\Theta_0$  or to the  $W$ -random graph model with graphon  $W_0$ . Then, given a single observation  $\mathbf{A}$ , we want to estimate  $\Theta_0$  or  $f_0$ .

Graphon estimation is more challenging than probability matrix estimation, in particular, because of identifiability issues: multiple graphons can lead to the same distribution on the space of graphs of size  $n$ . This is not unexpected as the distribution of the network is invariant with respect to any change of labeling of its nodes. More precisely, two graphons  $U$  and  $W$  in  $\mathcal{W}^+$  define the same probability distribution if and only if there exist measure preserving maps  $\phi, \psi: [0, 1] \rightarrow [0, 1]$  such that  $U(\phi(x), \phi(y)) = W(\psi(x), \psi(y))$  almost everywhere. This equivalence relation is called a weak isomorphism [28]. The corresponding quotient space is denoted by  $\widetilde{\mathcal{W}}^+$ . As a consequence, one can only estimate the equivalence class of  $\rho_n W_0$  in  $\widetilde{\mathcal{W}}^+$  and we refer henceforth to graphon estimation as the problem of estimating this equivalence class from the adjacency matrix  $\mathbf{A}$  sampled from the  $W$ -random graph model (1). When there is no ambiguity, we shall identify a graphon  $W \in \mathcal{W}^+$  and its corresponding equivalence class.

The problem of estimating  $\Theta_0$  was previously considered in a number of papers. For matrix estimation problem, the quality of an estimator  $\hat{\Theta}$  is usually assessed through the Frobenius loss  $\|\hat{\Theta} - \Theta_0\|_2$ . For instance, [16] obtain sub-optimal convergence rates for this problem using a singular thresholding algorithm. Relying on a least-square estimator [20] have established the minimax estimation rates for  $\Theta_0$  on classes of block constant matrices and smooth graphon classes. Their analysis is restricted to the dense case with constant  $\|\Theta_0\|_\infty$ . More recently, [26] extended their results to sparse case when  $\|\Theta_0\|_\infty$  depends on  $n$  and goes to zero when  $n \rightarrow \infty$ .

As for graphon estimation, most of results on estimation error are expressed in terms of  $l_2$  loss  $\|\widehat{W} - W_0\|_2$  (see below for a formal definition of this metric). For classes of smooth graphons, estimators based on maximum likelihood, restricted least-squares estimators, or neighborhood smoothing have been studied in [1, 14, 15, 26, 33, 35]. For classes of step-function graphons, restricted least-squares estimators have been considered in [10, 26] and the minimax optimal rates of convergence have been derived in [26].

Although one can take advantage of the Euclidean structure of the Frobenius matrix norm and the  $l_2$  metric on  $\mathcal{W}^+$ , both these metrics do not readily reflect the closeness in terms of the topology of the random graphs. As the structure of the graphon space is infinite-dimensional, not all norms are equivalent and one may wonder whether one should not study the graphon estimation problem with respect to a more suitable distance. We argue below that the cut distance which plays a central role in the random graph theory is a good candidate for this.

## 1.1 Cut metric

One of the fundamental questions in graph theory is the following one: what does it mean for two large graphs to be similar or close? There are different ways of defining the distance of two graphs. For example, the edit distance is defined as normalized Hamming distance of the edge

sets. Up to a normalization, it corresponds to  $l_1$  distance between the adjacency matrices. One of the troubles with this notion of distance is that it does not reflect well structural similarities between two graphs. For instance, the edit distance between two independent graphs drawn from the Erdős-Rényi model  $\mathcal{G}(n, p)$  with  $p = 1/2$  is close to  $1/2$  with high probability. Another notion of distance, called *cut distance*, better reflects the structural similarity. The cut norm of a matrix  $\mathbf{B} = (B_{ij}) \in \mathbb{R}^{n \times n}$  has been introduced by Frieze and Kannan [19]. It is defined by

$$\|\mathbf{B}\|_{\square} = \frac{1}{n^2} \max_{S, T \subseteq [n]} \left| \sum_{i \in S, j \in T} B_{ij} \right|.$$

In other words,  $\|\mathbf{B}\|_{\square}$  corresponds (up to a renormalization) to the maximal sum of entries over all submatrices of  $\mathbf{B}$ . Then, the cut distance  $d_{\square}(G, G')$  between two graphs  $G$  and  $G'$  defined on the same set of nodes and with adjacency matrices  $\mathbf{A}$  and  $\mathbf{A}'$  is defined as the cut norm  $\|\mathbf{A} - \mathbf{A}'\|_{\square}$ . Denoting  $e_G(S, T)$  the number of edge between nodes in  $S$  and  $T$  in the graph  $G$ , the cut distance  $d(G, G')$  is the supremum over all  $S, T$  of  $(e_G(S, T) - e_{G'}(S, T))/n^2$ . In other words,  $d_{\square}(G, G')$  is small if the restrictions of  $G$  and  $G'$  to all subsets  $S, T$  have similar edge densities.

Let us denote  $\mathcal{W}$  the collection of symmetric measurable functions  $[0, 1]^2 \rightarrow [-1, 1]$ . By analogy with the matrix cut norm, we can define the cut norm of a kernel  $W \in \mathcal{W}$ :

$$\|W\|_{\square} = \sup_{S, T \subseteq [0, 1]} \left| \int_{S \times T} W(x, y) dx dy \right|, \quad (2)$$

where the supremum is taken over all measurable subsets  $S$  and  $T$ . Then, the distance  $d_{\square}(W, W')$  between two graphons  $W$  and  $W'$  in  $\mathcal{W}^+$  is simply  $\|W - W'\|_{\square}$ . As explained earlier in the introduction, graphons in  $\mathcal{W}^+$  are not identifiable. This is why we consider the metric induced by  $\|\cdot\|_{\square}$  on the quotient space  $\widetilde{\mathcal{W}}^+$  defined by

$$\delta_{\square}(W_1, W_2) = \inf_{\tau \in \mathcal{M}} \|W_1 - W_2^{\tau}\|_{\square}, \quad (3)$$

where we take the infimum in the set  $\mathcal{M}$  of all measure-preserving bijections  $\tau : [0, 1] \rightarrow [0, 1]$  and  $W^{\tau}(x, y) = W(\tau(x), \tau(y))$ .

The cut distance is also a cornerstone in the graph limit theory introduced by Lovász and Szegedy [29] and further developed in, e.g., [8, 9]. In particular, this theory states that graphons can be interpreted as limits (with respect to  $\delta_{\square}$ ) of graph sequences. Besides, convergence in  $\delta_{\square}$  is equivalent to other structural properties such as the convergence of all homomorphism numbers. Given a simple graph  $F$  with  $q$  nodes and a graphon  $W_0$ , the homomorphism number  $t(F, W_0)$  is the probability that the edge set of size  $q$  of a graph sampled from the model (1) (with  $\rho_n = 1$ ) contains the edge set of  $F$ . As a consequence, the homomorphism numbers  $t(F, W_0)$  and  $t(F, W'_0)$  are close when the expected number of subgraphs  $F$  for a size  $n$  random graph  $G$  sampled from  $W_0$  is close to that of a size  $n$  random graph sampled from  $W'_0$ . It has been established that convergence in the cut distance is equivalent to convergence of homomorphism numbers for all simple graphs  $F$  (see Theorem 11.5 in [28] for more details). Hence, estimating well the graphon  $W_0$  in the cut distance allows to estimate well the number of small patterns induced by  $W_0$ . On the other hand, the cut distance controls other quantities of interest for computer scientists such as the size of multi-way cuts [10, 12]. So, a consistent estimator of  $W_0$  in cut distance gives consistent estimators for the multi-way cuts.

The construction of  $\delta_{\square}$  can be extended to any other norm  $N$  that is invariant under measure preserving maps:

$$\delta_N(W_1, W_2) = \inf_{\tau \in \mathcal{M}} \|W_1 - W_2^{\tau}\|_N. \quad (4)$$

Besides the cut norm, we already mentioned the  $l_1$  and  $l_2$ -norms on  $\mathcal{W}$  defined by  $\|W\|_1 = \int_{[0,1]^2} |W(x,y)| dx dy$  and  $\|W\|_2 = [\int_{[0,1]^2} W^2(x,y) dx dy]^{1/2}$ . These two norms define the corresponding distances  $\delta_1$  and  $\delta_2$  on the quotient space  $\widetilde{\mathcal{W}}^+$ . The distance  $\delta_\square$  is dominated by  $\delta_1$  and  $\delta_2$  (for details see Section 2.2). As already noted for instance in [10], this immediately implies that the convergence rate of an estimator  $\widetilde{W}$  with respect to the  $\delta_\square$ -distance is at least as fast as its convergence rate with respect to the  $\delta_2$ -distance. Then, one may wonder whether the convergence rates in  $\delta_\square$ -distance can be significantly faster and whether those faster rates are achieved by the estimators that are already minimax optimal with respect to other metrics.

In fact, a partial result on uniform convergence rates has already been proved. One of the striking consequences of the celebrated Szemerédi’s Lemma [31] states that an adjacency matrix sampled from a  $W$ -random graph model converges to the true graphon  $W_0$  in cut distance, this at an *uniform* rate over all graphons. To be more specific, let  $W_0 \in \mathcal{W}^+$  be a graphon and let  $\mathbf{A}$  be the size  $n$  adjacency matrix sampled according to the  $W$ -random graph model (1) with  $\rho_n = 1$ . It has been shown in [8] (see also [2] or [28]) that, with high probability, the empirical graphon  $\tilde{f}_{\mathbf{A}}$  associated to the adjacency matrix  $\mathbf{A}$  (see (18) for a precise definition) is  $O(1/\sqrt{\log(n)})$  close in the cut distance to the true graphon  $W_0$ :

**Proposition 1** (Lemma 10.16 [28]). *Let  $n \geq 1$  and let  $W_0 \in \mathcal{W}^+$  be a graphon. Then, with probability at least  $1 - \exp\{-n/(2 \log n)\}$ ,*

$$\delta_\square(\tilde{f}_{\mathbf{A}}, W_0) \leq \frac{22}{\sqrt{\log(n)}}. \quad (5)$$

An important point is that the above result is valid for all  $W_0 \in \mathcal{W}^+$ . Note that if we replace the cut-distance by  $\delta_1$  or  $\delta_2$ -distance this is not true any more: even in the simple case of a constant graphon  $W_0 \equiv a$  (with  $a \in (0, 1)$ ), the  $l_2$  distance between  $\tilde{f}_{\mathbf{A}}$  and  $W_0$  does not converge to zero.

## 1.2 Our contribution and related results

Our purpose in this paper is to go beyond uniform convergence rates over all graphons in  $\mathcal{W}^+$  and to understand the optimal cut distance convergence rates when  $W_0$  has a specific structure. First, optimal convergence rates are derived for the estimation of the connection probability matrix  $\Theta_0$  when it belongs to classes of block-constant matrices. Second, we establish the optimal convergence rates for all classes of step-function graphons  $f = \rho_n W_0$  both in sparse and dense case. In particular for  $\rho_n = 1$  (dense case), our results imply that, for any integer  $k \in [2, n]$  and  $k$ -steps graphon  $W_0$ , one has

$$\mathbb{E}_{W_0} [\delta_\square(\tilde{f}_{\mathbf{A}}, W_0)] \leq C \sqrt{\frac{k}{n \log(k)}}, \quad (6)$$

where  $C$  is a numerical constant (independent of  $n$  and  $k$ ) and that this convergence rate is optimal from the minimax point of view. This result has some interesting implications. In particular, this guarantees the optimality of the  $\log(n)^{-1/2}$  rate in Proposition 1 for general graphons. On the other hand, our results imply that for more structured classes of graphons ( $k \ll n$ ) much faster rates are achievable. Interestingly, we show that the adjacency matrix and its associated empirical graphons are already adaptive to the unknown number of blocks of the matrix  $\Theta_0$  or steps of  $W_0$  and minimax optimal. As a consequence, there is no need to look for more involved estimators.

In practice, it could be disappointing that the raw data are already optimal with respect to the cut distance, whereas they perform really badly with respect to the  $\delta_2$  distance. This is why we prove that a singular value hard thresholding estimator is still optimal with respect to the

cut metric  $\delta_{\square}$  while achieving the best known rate in  $\delta_2$ -distance in the class of polynomial-time estimators.

Our results are in sharp contrast to all aforementioned manuscripts [1, 10, 14, 15, 26, 33, 35] whose primary focus is the  $\delta_2$ -distance and whose convergence rates with respect to the  $\delta_{\square}$ -distance are derived from the domination of  $\delta_{\square}$  by  $\delta_2$ . Closest to our contributions, is the recent paper [7] where the authors introduce a least-cut norm estimator for a more general model of unbounded graphons. Translated in our framework, their non-polynomial time algorithm achieves, in some cases, the optimal convergence rate (up to a logarithmic loss) and it is slower in other cases. In Section 4.3 we extend our study to unbounded graphons and compare our results to those of [7]. In particular, our Proposition 7 implies that the empirical graphon associated to the adjacency matrix and to the singular value hard thresholding estimator are optimal (up to a logarithmic factor) also in the general case of unbounded graphons. Note that the main difference with the method proposed in [7] is that both our estimators can be easily computed in polynomial time.

From a technical point of view, the tools needed for deriving optimal cut distance rates differ from those used for the  $\delta_2$ -distance. For establishing the minimax lower bounds, the main technical hurdle is to build a collection of well-spaces graphons with respect to the cut distance. Indeed, the cut distance  $\delta_{\square}(W_1, W_2)$  is difficult to lower bound as it is defined as an infimum over all measure-preserving transformations. As for the minimax upper bound on the estimation error in (6), it can be obtained quite easily without the correct logarithmic term thanks to the Bernstein inequality together with some bounds from [26] for the stronger metric  $\delta_2$ . However, recovering the right logarithmic term in (6) is much more challenging. The proof relies among other things on a careful application of Szemerédi’s regularity lemma to distorted versions of the graphon.

The manuscript is organized as follows. First, we recall some basic results related to the cut metric. The problem of estimating the matrix of connection probabilities is considered in Section 3. We study the problem of graphon estimation in Section 4. The appendix contains all the proofs where in Appendix A we recall some basic facts and results that are often used in the proofs.

## 2 Notation and Preliminaries

### 2.1 Notation

We gather here some of the notation used throughout this paper. Some of them have already been defined in the introduction.

- For a matrix  $\mathbf{B}$ ,  $\mathbf{B}_{ij}$  (or  $\mathbf{B}_{i,j}$ , or  $(\mathbf{B})_{ij}$ ) is its  $(i, j)$ -th entry. Let  $\mathbf{B}_{i,\cdot}$  and  $\mathbf{B}_{\cdot,j}$  stand for its  $i$ th row and  $j$ th column respectively. We denote by  $\mathbb{R}_{\text{sym}}^{k \times k}$  the class of all symmetric  $k \times k$  matrices with real-valued entries. Given a matrix  $\mathbf{B}$  and  $p \in [1, \infty]$ ,  $\|\mathbf{B}\|_p$  denotes its entry-wise  $l_p$  norm, that is  $\|\mathbf{B}\|_p^p = \sum_{i,j} |\mathbf{B}_{ij}|^p$  for  $p < \infty$  and  $\|\mathbf{B}\|_{\infty} = \max_{i,j} |\mathbf{B}_{ij}|$ . Given  $(p, q) \in [1, \infty]$ ,  $\|\mathbf{B}\|_{p \rightarrow q}$  stands for its  $l_p \rightarrow l_q$  operator norm. Finally,  $\langle \mathbf{D}, \mathbf{B} \rangle = \sum_{i,j} \mathbf{D}_{ij} \mathbf{B}_{ij}$  stands for the canonical inner product between matrices  $\mathbf{D}, \mathbf{B} \in \mathbb{R}^{n \times n}$ .
- $\mathcal{W}$  is the collection of symmetric measurable functions  $[0, 1]^2 \rightarrow [-1, 1]$ . Given a kernel  $W \in \mathcal{W}$  and  $p \in (1, \infty)$ , its  $l_p$  norm is defined by  $\|W\|_p^p = \int |W(x, y)|^p dx dy$ , whereas  $\|W\|_{\infty} = \text{ess sup}_{x,y} |W(x, y)|$ .  $\mathcal{W}^+$  is the space of graphons and  $\widetilde{\mathcal{W}}^+$  is the corresponding quotient space. The cut distance  $\delta_{\square}(\cdot, \cdot)$  in the graphon spaces is defined by (3). Also,  $\delta_1(\cdot, \cdot)$  and  $\delta_2(\cdot, \cdot)$  defined by (4) respectively correspond to the  $l_1$  and  $l_2$  distances on the quotient space of graphons  $\widetilde{\mathcal{W}}^+$ . Given a symmetric square matrix  $\Theta$  with values in  $[0, 1]$ ,  $\widehat{f}_{\Theta}$  is the empirical graphon  $\Theta$  as defined in (18).

- Given a probability matrix  $\Theta_0$ , we denote by  $\mathbb{E}_{\Theta_0}$  the expectation with respect to the distribution of  $\mathbf{A}$  if we consider the inhomogeneous random graph model and given a graphon  $W$  and  $\rho_n$ , we write  $\mathbb{E}_W$  for the expectation with respect to the joint distribution of  $(\xi, \mathbf{A})$ .
- We denote by  $\lfloor x \rfloor$  the maximal integer less than or equal to  $x$  and by  $\lceil x \rceil$  the smallest integer greater than or equal to  $x$ . For an positive integer  $m$ , set  $[m] = \{1, \dots, m\}$ .  $\mathbb{1}_A(\cdot)$  denotes the indicator function of a set  $A$ . In the sequence,  $C$  stands for a positive constant that can vary from line to line. These are absolute constants unless otherwise mentioned. For two positive functions  $f$  and  $g$ , we write  $f \asymp g$  when there exist two positive numerical constants  $C$  and  $C'$  such  $Cg \leq f \leq C'g$ . Finally,  $\lambda$  is the Lebesgue measure on the interval  $[0, 1]$ .

## 2.2 Preliminaries

We start with a few basic properties of the cut norm for matrices  $\mathbf{A}$  and graphons  $W$ . It is easy to see that

$$\|\mathbf{A}\|_{\square} \leq \frac{1}{n^2} \|\mathbf{A}\|_1 \leq \frac{1}{n} \|\mathbf{A}\|_2$$

where  $\|\cdot\|_1$  and  $\|\cdot\|_2$  are the usual entry-wise  $l_1$  and  $l_2$ -norms of a matrix. For a function  $W \in \mathcal{W}$ , we have

$$\|W\|_{\square} \leq \|W\|_1 \leq \|W\|_2 \leq \|W\|_{\infty} \leq 1$$

where  $\|\cdot\|_1$  and  $\|\cdot\|_2$  denote  $l_1$  and  $l_2$ -norms of a graphon. In the opposite direction, we have  $\|W\|_2 \leq \sqrt{\|W\|_1}$ . As a consequence, the metric  $\delta_1$  and  $\delta_2$  define the same topology on the space  $\widetilde{\mathcal{W}}^+$  of graphons. In contrast, the cut distance  $\delta_{\square}$  defines a weaker topology on the space  $\widetilde{\mathcal{W}}^+$  as illustrated by the aforementioned sampling result (Proposition 1).

We shall also sometimes rely on the equivalence between the cut norm and to the  $l_{\infty} \rightarrow l_1$  operator norm:

$$\|W\|_{\infty \rightarrow 1} = \sup_{\|f\|_{\infty}, \|g\|_{\infty} \leq 1} \left| \int_{S \times T} W(x, y) f(x) g(y) dx dy \right| \quad (7)$$

where the supremum is taken over all (real-valued) functions  $f$  and  $g$  with values in  $[-1, 1]$ . It is known that (see e.g., [24])

$$\|W\|_{\square} \leq \|W\|_{\infty \rightarrow 1} \leq 4\|W\|_{\square}. \quad (8)$$

## 3 Probability matrix estimation

### 3.1 Cut norm minimax risk

We start with a simple proposition that bounds the expected cut distance between  $\Theta_0$  and the sampled adjacency matrix  $\mathbf{A}$ . Similar results already appeared in the literature, see e.g., [28, Lemma 10.11], [7] or [21]. Its proofs is based on Bernstein inequality and is given in Section B.

**Proposition 2.** *For any probability matrix  $\Theta_0$  we have*

$$\mathbb{E}_{\Theta_0} \|\mathbf{A} - \Theta_0\|_{\square} \leq 12 \sqrt{\frac{\|\Theta_0\|_1 + n}{n^3}}. \quad (9)$$

*In particular, if  $\|\Theta_0\|_{\infty} \geq 1/n$ , we get*

$$\mathbb{E}_{\Theta_0} \|\mathbf{A} - \Theta_0\|_{\square} \leq 24 \sqrt{\frac{\|\Theta_0\|_{\infty}}{n}}.$$



This implies that the adjacency matrix  $\mathbf{A}$  is  $\sqrt{\|\boldsymbol{\Theta}_0\|_\infty/n}$ -close in cut-distance to the probability matrix  $\boldsymbol{\Theta}_0$ . This bound is valid for all matrices  $\boldsymbol{\Theta}_0$ . It turns out that no estimator can perform much better than  $\mathbf{A}$ , even on some simple classes of parameters  $\boldsymbol{\Theta}_0$ .

Let  $n, k$  be integers such that  $2 \leq k \leq n$  and  $\mathcal{T}[k]$  be defined by

$$\mathcal{T}[k] = \{\boldsymbol{\Theta}_0 : \exists z \in \mathcal{Z}_{n,k}, \mathbf{Q} \in [0, 1]_{\text{sym}}^{k \times k} \text{ such that } \boldsymbol{\Theta}_{ij} = \mathbf{Q}_{z(i)z(j)}, i \neq j, \text{ and } \boldsymbol{\Theta}_{ii} = 0 \forall i\}$$

where we denote by  $\mathcal{Z}_{n,k}$  the set of all mappings  $z$  from  $[n]$  to  $[k]$ . In other words  $\mathcal{T}[k]$  is made of matrices that, up to a permutation of their rows and their columns, are (up to the diagonal) block constants with at most  $k$  blocks. Also, this corresponds to connection probability matrices of  $k$ -class stochastic blocks models whose vector label  $Z = (Z_a)$  has been fixed. For any  $\rho_n \in (0, 1]$ , consider the set

$$\mathcal{T}[k, \rho_n] = \{\boldsymbol{\Theta}_0 \in \mathcal{T}[k] : \|\boldsymbol{\Theta}_0\|_\infty \leq \rho_n\},$$

of matrices whose largest value is smaller or equal to  $\rho_n$ . The following Proposition, proved in section C, gives a lower bound on the minimax risk over the class  $\mathcal{T}[2, \rho_n]$  of block-constant matrices with only two blocks:

**Proposition 3.** *The minimax risk measured in cut norm satisfies*

$$\inf_{\hat{\boldsymbol{\Theta}}} \sup_{\boldsymbol{\Theta}_0 \in \mathcal{T}[2, \rho_n]} \mathbb{E}_{\boldsymbol{\Theta}_0} [\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_\square] \geq C \min \left( \sqrt{\frac{\rho_n}{n}}, \rho_n \right)$$

where  $\mathbb{E}_{\boldsymbol{\Theta}_0}$  denotes the expectation with respect to the distribution of  $\mathbf{A}$  when the underlying probability matrix is  $\boldsymbol{\Theta}_0$ .

Comparing Proposition 3 with Proposition 2 we observe that the raw data  $\mathbf{A}$  is minimax optimal for the class  $\mathcal{T}[2, \rho_n]$  for all  $\rho_n \geq 1/n$ . As a consequence, there is no need to look for a more involved estimator. Since for  $\rho_n \leq 1/n$  the constant estimator  $\hat{\boldsymbol{\Theta}} = 0$  satisfies  $\mathbb{E}_{\boldsymbol{\Theta}_0} [\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_\square] \leq \rho_n$  and using that the collections  $\mathcal{T}[k, \rho_n]$  are nested, the two previous propositions imply that the optimal cut norm estimation rates for  $\mathcal{T}[k, \rho_n]$  with  $k \geq 2$  is given by

$$\inf_{\hat{\boldsymbol{\Theta}}} \sup_{\boldsymbol{\Theta}_0 \in \mathcal{T}[k, \rho_n]} \mathbb{E}_{\boldsymbol{\Theta}_0} [\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_\square] \asymp \min \left( \sqrt{\frac{\rho_n}{n}}, \rho_n \right).$$

Until now, we left aside the specific case of constant matrices  $\mathcal{T}[1, \rho_n]$  which correspond to Erdős-Renyi random graphs. It turns out that the situation is quite different for this simple class. For a constant matrix  $\boldsymbol{\Theta}_0$ , estimating  $\boldsymbol{\Theta}_0$  given  $\mathbf{A}$  amounts to infer the parameter  $p$  of a Bernoulli distribution given a sample of size  $n(n-1)/2$ . From this analogy, we consider the matrix  $\overline{\mathbf{A}}$  whose all non-diagonal entries are equal to  $\sum_{i,j} \mathbf{A}_{ij} / (n(n-1))$ . Then, it is straightforward to prove that

$$\mathbb{E}_{\boldsymbol{\Theta}_0} [\|\boldsymbol{\Theta}_0 - \overline{\mathbf{A}}\|_\square] \leq \sqrt{\frac{2\rho_n}{n(n-1)}},$$

which is  $\sqrt{n}$ -faster than what is achieved by the adjacency matrix  $\mathbf{A}$ . Using again the analogy with the problem of Bernoulli parameter estimation, one may easily get the following minimax lower bound:

$$\inf_{\hat{\boldsymbol{\Theta}}} \sup_{\boldsymbol{\Theta}_0 \in \mathcal{T}[1, \rho_n]} \mathbb{E}_{\boldsymbol{\Theta}_0} [\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_\square] \geq C \min \left( \frac{\sqrt{\rho_n}}{n}, \rho_n \right)$$

which assesses that the  $\sqrt{\rho_n}/n$ -rate achieved by  $\overline{\mathbf{A}}$  is optimal.

### 3.2 Comparison with $l_1$ and $l_2$ -estimation

The cut norm optimal estimation rate is quite different from what has been established for the Frobenius norm (also called  $l_2$ ) estimation rate in [26] (see also [20] for the dense case), that is

$$\inf_{\hat{\Theta}} \sup_{\Theta_0 \in \mathcal{T}[k, \rho_n]} \mathbb{E}_{\Theta_0} \left[ \frac{1}{n} \left\| \hat{\Theta} - \Theta_0 \right\|_2 \right] \asymp \min \left( \sqrt{\frac{\rho_n \log(k)}{n}} + \frac{\sqrt{\rho_n k}}{n}, \rho_n \right), \quad (10)$$

for any  $k = 2, \dots, n$ . Besides, the minimax risk bound is achieved by the restricted least-square estimators [26] defined by

$$\hat{\Theta}_{k, \rho_n} := \arg \min_{\Theta \in \mathcal{T}[k, \rho_n]} \|\Theta - \mathbf{A}\|_2^2. \quad (11)$$

Since the Frobenius norm dominates the cut norm, it is expected that the cut norm convergence rate is faster than the Frobenius norm estimation rate. When  $\rho_n$  is not too small and the number of blocks remains small ( $k \leq \sqrt{n \log(n)}$ ), the gain is a  $\log(k)$  factor, whereas, for larger  $k$ , the gain is of order  $k/\sqrt{n}$ . More importantly, the optimal Frobenius norm convergence rate (10) is only known to be achieved by non-polynomial time estimators such as (11).

In view of the above discussion, one may wonder whether it is possible to build estimators that are near optimal in terms of both the cut and Frobenius distances. Since for any matrix  $\mathbf{B}$ ,  $\|\mathbf{B}\|_{\square} \leq \|\mathbf{B}\|_2/n$ , it follows that, for  $k \leq \sqrt{n}$ , the restricted least-square estimator  $\hat{\Theta}_{k, \rho_n}$  (11) is also near optimal (up to  $\sqrt{\log(k)}$  factor) with respect to the cut distance, that is,

$$\mathbb{E}_{\Theta_0} \left[ \left\| \hat{\Theta}_{k, \rho_n} - \Theta_0 \right\|_{\square} \right] \leq C \sqrt{\frac{\rho_n \log(k)}{n}}.$$

For matrices  $\Theta_0$  with more than  $\sqrt{n}$  blocks, it is not clear whether the estimator  $\hat{\Theta}_{k, \rho_n}$  achieves a fast rate of convergence in the cut norm.

In any case, the computational complexity of  $\hat{\Theta}_{k, \rho_n}$  is non polynomial. In fact, no polynomial-time algorithm is known to achieve the minimax risk (10) with respect to the Frobenius norm. Below, we describe an estimator that is optimal in the cut distance and also achieves the best known rate in Frobenius distance in the class of polynomial-time estimators. Let us write the singular value decomposition of  $\mathbf{A}$ :

$$\mathbf{A} = \sum_{j=1}^{\text{rank}(\mathbf{A})} \sigma_j(\mathbf{A}) u_j(\mathbf{A}) v_j(\mathbf{A})^T, \quad (12)$$

where  $\sigma_j(\mathbf{A}) > 0$  are the singular values of  $\mathbf{A}$  indexed in the decreasing order,  $u_j(\mathbf{A})$  are eigenvectors of  $\mathbf{A}$  and  $v_j(\mathbf{A}) = \pm u_j(\mathbf{A})$ . Given a tuning parameter  $\lambda > 0$ , we define

$$\tilde{\Theta}_{\lambda} = \sum_{j: \sigma_j(\mathbf{A}) \geq \lambda} \sigma_j(\mathbf{A}) u_j(\mathbf{A}) v_j(\mathbf{A})^T \quad (13)$$

as the singular value hard thresholding estimator of  $\Theta_0$ . We have the following

**Proposition 4.** *Assume that  $\rho_n \geq \log(n)/n$ . Let  $\lambda = c\sqrt{\rho_n n}$  where  $c$  is a sufficiently large numerical constant. Then, for any  $k \in [n]$  and any  $\Theta_0 \in \mathcal{T}[k, \rho_n]$ , the hard thresholding estimator  $\tilde{\Theta}_{\lambda}$  simultaneously satisfies, with probability larger than  $1 - 1/n$ ,*

$$\frac{1}{n} \|\tilde{\Theta}_{\lambda} - \Theta_0\|_2 \leq C \sqrt{\frac{\rho_n k}{n}}, \quad (14)$$

$$\|\tilde{\Theta}_{\lambda} - \Theta_0\|_{\square} \leq \frac{1}{n} \|\tilde{\Theta}_{\lambda} - \Theta_0\|_{2 \rightarrow 2} \leq C \sqrt{\frac{\rho_n}{n}}, \quad (15)$$

where  $C$  is a numerical constant.



The low-rank estimator  $\tilde{\Theta}_\lambda$  was previously considered in [16] for Frobenius norm estimation, but error bounds obtained in [16] are much more pessimistic than (14). It follows from (15), that for  $\rho_n \geq \log(n)/n$ , with high probability,  $\tilde{\Theta}_\lambda$  achieves the optimal rate in the cut norm and the  $\sqrt{\rho_n k}/n$  rate in Frobenius norm, which is the best known rate among polynomial-time estimators.

We close this section by the following proposition which gives the minimax optimal rate of estimation in  $l_1$ -norm. This will allow us to further compare the  $\delta_1$  and  $\delta_\square$  convergence rates for graphon estimation in the next section.

**Proposition 5.** *For any sequence  $\rho_n > 0$  and any positive integer  $2 \leq k \leq n$ , one has*

$$\inf_{\hat{\Theta}} \sup_{\Theta_0 \in \mathcal{T}[k, \rho_n]} \mathbb{E}_{\Theta_0} \left[ \frac{1}{n^2} \left\| \hat{\Theta} - \Theta_0 \right\|_1 \right] \asymp \min \left\{ \sqrt{\frac{\rho_n \log(k)}{n}} + \frac{\sqrt{\rho_n k}}{n}, \rho_n \right\}. \quad (16)$$

The upper bound in (16) is a consequence of the inequality  $\|\mathbf{B}\|_1/n^2 \leq \|\mathbf{B}\|_2/n$  together with the control of the estimation error of the restricted least-squares estimator  $\hat{\Theta}_{k, \rho_n}$  (11) performed in [26]. The lower bound of the minimax risk in (16) is proved following the same lines as the proof of Proposition 2.4 in [26] with  $\|\cdot\|_2$  replaced by  $\|\cdot\|_1$ . We skip the details.

## 4 Graphon estimation problem

In this section, we are interested in estimating the graphon  $W_0$  in the sparse  $W$ -random graph model (1). Let  $\mathcal{W}^+[k]$  be the collection of  $k$ -step graphons, that is, the subset of graphons  $W \in \mathcal{W}^+$  such that for some  $\mathbf{Q} \in [0, 1]_{\text{sym}}^{k \times k}$  and some  $\phi : [0, 1] \rightarrow [k]$ ,

$$W(x, y) = \mathbf{Q}_{\phi(x), \phi(y)} \quad \text{for all } x, y \in [0, 1]. \quad (17)$$

Note  $\mathcal{W}^+[k]$  is also in correspondence with the collection of stochastic block models with  $k$  blocks. Our purpose here, is to characterize the minimax convergence rates over classes  $\mathcal{W}^+[k]$ .

### 4.1 Cut distance minimax risk

Following [26], we start by associating a graphon to any  $n \times n$  probability matrix  $\Theta_0$ . Then, we can estimate graphon  $f_0(\cdot, \cdot) = \rho_n W_0(\cdot, \cdot)$  using the empirical graphon associated to an estimator of  $\Theta_0$ . Given a  $n \times n$  matrix  $\Theta$  with entries in  $[0, 1]$ , we define the graphon  $\tilde{f}_\Theta$  as the following piecewise constant function:

$$\tilde{f}_\Theta(x, y) = \Theta_{[nx], [ny]} \quad (18)$$

for all  $x$  and  $y$  in  $(0, 1]$ . For any estimator  $\hat{T}$  of  $\Theta_0$  and any norm  $N$  that is invariant under measure preserving maps the triangle inequality implies

$$\mathbb{E}_{W_0} \left[ \delta_N(\tilde{f}_{\hat{T}}, f_0) \right] \leq \mathbb{E}_{W_0} \left[ \|\hat{T} - \Theta_0\|_N \right] + \mathbb{E}_{W_0} \left[ \delta_N(\tilde{f}_{\Theta_0}, f_0) \right]. \quad (19)$$

We have two parts in (19). The first term is the *estimation error* term  $\|\hat{T} - \Theta_0\|_N$  that has been considered in the previous section. The second term  $\delta_N(\tilde{f}_{\Theta_0}, f_0)$  is the *agnostic error*. It measures the  $\delta_N$ -distance between the true graphon  $f_0$  and its discretized version sampled at the unobserved random design points  $\xi_1, \dots, \xi_n$ . The behavior of  $\delta_N(\tilde{f}_{\Theta_0}, f_0)$  depends on the topology of the considered class of graphons. The following theorem, proved in Section E, gives the upper bound on the agnostic error, measured in  $\delta_\square$ -distance for step function graphons:

**Theorem 1** (Agnostic error measured in cut distance). *Consider the  $W$ -random graph model (1). For all integers  $k \geq 2$ , all positive integers  $n$ , all  $W_0 \in \mathcal{W}^+[k]$  and  $\rho_n > 0$ , we have*

$$\mathbb{E}_{W_0} \left[ \delta_{\square}(\tilde{f}_{\Theta_0}, f_0) \right] \leq C \rho_n \begin{cases} \sqrt{\frac{k}{n \log(k)}} & \text{if } k \leq n, \\ \sqrt{\frac{1}{\log(n)}} & \text{if } k > n. \end{cases}$$

Note that the case  $k > n$  is a consequence of Proposition 1 from [28], so that we effectively only have to consider the case  $k \leq n$ . The proof combines two ideas. First, we build  $W$  and  $\widehat{W}$  as the representatives of  $W_0$  and  $\tilde{f}_{\Theta_0}$  in the quotient space  $\widetilde{\mathcal{W}}^+$  such that  $W$  and  $\widehat{W}$  match everywhere except on a set of Lebesgue measure of order at most  $\sqrt{k/n}$ . This allows us to get a risk bound of order  $\sqrt{k/n}$ . In order to recover the correct logarithmic factor  $\sqrt{\log(k)}$ , we rely on the weak Szemerédy Lemma. Here, the key idea is to build a cut-norm approximation of a distorted transformation of  $W$  where the weights of the group have been modified to take into account the geometry of the sampling error.

As an immediate consequence of (19), Proposition 2 and Theorem 1, we get the following upper bound on the risk of the empirical graphon  $\tilde{f}_{\mathbf{A}}$ . For any  $k \geq 2$ , it holds that

$$\sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0} \left[ \delta_{\square}(\tilde{f}_{\mathbf{A}}, f_0) \right] \leq C \min \left( \rho_n \left( \sqrt{\frac{k}{n \log(k)}}, \frac{1}{\sqrt{\log(n)}} \right) + \sqrt{\frac{\rho_n}{n}} \right), \quad (20)$$

where  $C$  is an absolute constant. Here,  $\mathbb{E}_{W_0}$  denotes the expectation with respect to the distribution of observations  $\mathbf{A} = (\mathbf{A}_{ij}, 1 \leq j < i \leq n)$  when the underlying sparse graphon is  $f_0 = \rho_n W_0$ . The following Proposition provides a matching lower bound for  $2 \leq k \leq n$ .

**Theorem 2.** *There exists a universal constant  $C > 0$  such that for any sequence  $\rho_n > 0$  and any positive integer  $2 \leq k \leq n$ ,*

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0} \left[ \delta_{\square}(\hat{f}, f_0) \right] \geq C \min \left( \rho_n \sqrt{\frac{k}{n \log(k)}} + \sqrt{\frac{\rho_n}{n}}, \rho_n \right), \quad (21)$$

where  $\inf_{\hat{f}}$  is the infimum over all estimators.

Since the collections  $\mathcal{W}^+[k]$  are nested, it follows that for all  $k \geq n$ , one has

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0} \left[ \delta_{\square}(\hat{f}, f_0) \right] \geq C \min \left( \rho_n \sqrt{\frac{1}{\log(n)}} + \sqrt{\frac{\rho_n}{n}}, \rho_n \right).$$

In view of (20) and (21), we observe that, as long as,  $\rho_n \geq 1/n$ , the empirical graphon  $\tilde{f}_{\mathbf{A}}$  is minimax optimal over all classes  $\mathcal{W}^+[k]$ ,  $k \geq 2$ . For sparser graphs ( $\rho_n \leq 1/n$ ), the trivial estimator  $\hat{f} \equiv 0$  achieves the optimal rate  $\rho_n$ .

Note that there are two distinct regimes in the minimax convergence rate. When  $\rho_n \geq \log(k)/k$  (weakly sparse graphs or large number of groups), the agnostic error dominates and the minimax risk is of order  $\rho_n \sqrt{k/(n \log(k))}$ . For moderately sparse graphs or equivalently a small number of steps ( $n^{-1} \leq \rho_n \leq \log(k)/k$ ), the error arising from the probability matrix  $\Theta_0$  estimation dominates and the minimax risk is of order  $\sqrt{\rho_n/n}$ .

As in the previous section, we left aside the specific case of constant graphons  $\mathcal{W}^+[1]$ . Note that for a graphon  $W_0 \in \mathcal{W}^+[1]$  the agnostic error is always zero and the loss comes from the probability matrix estimation. Following the arguments of the previous section, we derive that the graphon  $\tilde{f}_{\mathbf{A}}$  converges to  $\rho_n W_0$  at the rate  $\sqrt{\rho_n}/n$  which is optimal as soon as  $\rho_n \geq 1/n^2$ .

## 4.2 Comparison with $\delta_1$ and $\delta_2$ -estimation

Minimax risk for graphon estimation in the  $\delta_2$ -distance was obtained in [26, Proposition 3.2] :

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0} [\delta_2(\hat{f}, f)] \asymp \min \left( \frac{\sqrt{\rho_n k}}{n} + \sqrt{\frac{\rho_n \log(k)}{n}} + \rho_n \left( \frac{k}{n} \right)^{1/4}, \rho_n \right) \quad (22)$$

The following proposition, proved in Section G, gives the minimax  $\delta_1$ -convergence rate:

**Proposition 6.** *For any sequence  $\rho_n > 0$  and any positive integer  $2 \leq k \leq n$ , we have*

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0} [\delta_1(\hat{f}, f_0)] \geq C_1 \min \left( \frac{\sqrt{\rho_n k}}{n} + \sqrt{\frac{\rho_n}{n}} + \rho_n \sqrt{\frac{k}{n}}, \rho_n \right). \quad (23)$$

Conversely, there exists an estimator  $\hat{f}$  based on the restricted least-squares estimator (11) such that

$$\sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0} [\delta_1(\hat{f}_{\hat{\Theta}_{k, \rho_n}}, f_0)] \leq C_2 \min \left( \rho_n \sqrt{\frac{k}{n}} + \frac{\sqrt{\rho_n k}}{n} + \sqrt{\frac{\rho_n \log(k)}{n}}, \rho_n \right). \quad (24)$$

The upper and lower bounds given by Proposition 6 match (up to a  $\sqrt{\log(k)}$  multiplicative term in one of the regimes). There are three regions in (24) for  $\delta_1$  graphon estimation. The first one corresponds to the case of weakly sparse graphs with  $\rho_n \geq k^{-1} \vee (k/n)$ . In this case, the agnostic error dominates and the optimal risk is of order  $\rho_n \sqrt{k/n}$ . For moderately sparse graphs with  $n^{-1} \vee (k/n)^2 \leq \rho_n \leq k^{-1} \vee (k/n)$ , the probability matrix estimation error dominates and the minimax rate is of order  $\sqrt{\rho_n/n} + \sqrt{\rho_n k/n}$  (up to a  $\log(k)$  multiplicative term). In the case of highly sparse graphs with  $\rho_n \leq n^{-1} \vee (k/n)^2 \vee (\frac{k}{n})^2$ , the minimax risk is  $\rho_n$  which corresponds to the risk of the null estimator  $\tilde{f} \equiv 0$ .

Let us compare the optimal convergence rates with respect to the  $\delta_1$  (24),  $\delta_2$  (22) and  $\delta_\square$  (21). Bearing in mind that  $\delta_2$  dominates  $\delta_1$ , which in turn dominates  $\delta_\square$ , one should not be surprised that optimal rates with respect to  $\delta_2$  are the slowest. When the number of steps  $k$  is less than  $\sqrt{n}$  or when the graph is weakly sparse ( $\rho_n \geq \sqrt{k/n}$ ), then the  $\delta_1$  and  $\delta_\square$  optimal rates only differ by a  $\log(k)$  multiplicative term. For larger  $k$  and sparser graph, the optimal  $\delta_1$ -risk can be  $k/\sqrt{n}$  larger than the  $\delta_\square$ -risk.

Following the discussion in Section 3.2, one may easily build graphon estimators performing well in all these three distances. For instance, the graphon  $\hat{f}_{\hat{\Theta}_{k, \rho_n}}$  based on the restricted-least-squares estimator is optimal with respect to  $\delta_2$  and  $\delta_1$  and near optimal (up to a possible  $\sqrt{\log(k)}$  loss) with respect to  $\delta_\square$  for  $k \leq \sqrt{n}$ . Besides, the graphon  $\hat{f}_{\hat{\Theta}_\lambda}$  based on the singular value thresholding estimator is optimal with respect to  $\delta_\square$  and achieves best known convergence rates with respect to  $\delta_1$  and  $\delta_2$  among polynomial time algorithms.

## 4.3 Cut distance estimation of $L_1$ and $L_2$ graphons

Until now we have restricted our attention to graphons  $W$  taking values in  $[0, 1]$ . As argued in [11, 12], in this case the empirical degree distribution of a graph sampled from the corresponding  $W$ -random graph model (1) is light. This contrasts with many practical situations, where the degree distribution is heavy tailed. To circumvent this limitation, Borgs et al [11, 12] introduce, for  $p \geq 1$ , the class  $\mathcal{W}_p^+$  of symmetric measurable functions  $W : [0, 1]^2 \rightarrow \mathbb{R}^+$  such that  $\int |W(x, y)|^p dx dy < \infty$ . This collection  $\mathcal{W}_p^+$  is referred as the collection of  $L_p$  graphons. We have the inclusions

$\mathcal{W}^+ \subset \mathcal{W}_p^+ \subset \mathcal{W}_{p'}^+$  for  $p > p' \geq 1$ . Given a graphon  $W_0 \in \mathcal{W}_p^+$  and a sparsity parameter  $1 \geq \rho_n > 0$ , the corresponding  $W$ -random graph model amounts to generating a graph with  $n$  vertices according to the random matrix  $\Theta_0$  sampled as follows

$$\Theta_{ij} = [\rho_n W_0(\xi_i, \xi_j)] \wedge 1, \quad \forall i \neq j \text{ and } \Theta_{ii} = 0, \quad (25)$$

where  $\xi_1, \dots, \xi_n$  are, as in (1), i.i.d. random variables uniformly distributed in  $[0, 1]$ . Note that since  $W_0$  is now unbounded, we have to take the minimum with 1 in (25). We write  $f'_0 = (\rho_n W_0) \wedge 1$ . Since  $W_0$  is now allowed to be unbounded, graphs sampled according to the model (25) may have power law degree distribution [11]. As in the introduction, we may extend the norms  $\|\cdot\|_\square$  and  $\|\cdot\|_q$  and the distances  $\delta_\square$  and  $\delta_q$  to any graphon  $W_0 \in \mathcal{W}_p^+$  with  $p \leq q$ . Also, we write  $\mathcal{W}_p^+$  for the quotient space of  $L_p$  graphons under weak isometry.

Let us also define the collection  $\mathcal{W}_p^+[k]$  of  $k$ -steps  $L_p$  graphons, that is the subsets of graphon  $W \in \mathcal{W}_p^+$  such that  $W(x, y) = \mathbf{Q}_{\phi(x), \phi(y)}$  for some  $\mathbf{Q} \in (\mathbb{R}^+)^{k \times k}_{\text{sym}}$  and some  $\phi : [0, 1] \rightarrow [k]$  (note that  $\mathcal{W}_p^+[k]$  does not depend on  $p$ ). For  $1 \geq \mu > 0$  we denote by  $\mathcal{W}_p^+[k, \mu]$  the subset of  $\mathcal{W}_p^+[k]$  of “balanced” step functions, that is,  $W \in \mathcal{W}_p^+[k, \mu]$  if  $\lambda(\phi^{-1}(a)) \geq \mu/k$  for all  $a \in [k]$ . This means that the size of each step is larger than  $\mu/k$ .

Without loss of generality we can consider normalized graphons, that is, we assume that  $\|W_0\|_1 = 1$ . The following proposition proved in Appendix H gives an oracle inequality for the risk of the empirical graphon associated to the adjacency matrix and to the singular value hard thresholding estimator:

**Proposition 7.** *Let  $\lambda = c\sqrt{\rho_n n}$  where  $c$  is a sufficiently large numerical constant. Given a graphon  $W_0$  and  $\rho_n > 0$ , write  $W'_0 = \rho_n^{-1}[(\rho_n W_0) \wedge 1]$ .*

- (1) *Let  $W_0 \in \mathcal{W}_1^+$  with  $\|W_0\|_1 = 1$ ,  $\rho_n \geq 1/n$  and  $1 \geq \mu > 0$ . Then, for any positive integer  $k \leq \mu n$ , we have*

$$\mathbb{E}_{W_0} [\delta_\square(\tilde{f}_A, f'_0)] \leq 2\rho_n \inf_{W \in \mathcal{W}_1^+[k, \mu]} \delta_1(W, W'_0) + C \left[ \rho_n \sqrt{\frac{k}{\mu n}} + \sqrt{\frac{\rho_n}{n}} \right] \quad (26)$$

and

$$\mathbb{E}_{W_0} [\delta_\square(\tilde{f}_{\tilde{\Theta}_\lambda}, f'_0)] \leq 2\rho_n \inf_{W \in \mathcal{W}_1^+[k, \mu]} \delta_1(W, W'_0) + C \left[ \rho_n \sqrt{\frac{k}{\mu n}} + \sqrt{\frac{\rho_n}{n}} \right]. \quad (27)$$

- (2) *Assume that  $W_0 \in \mathcal{W}_2^+$  with  $\|W_0\|_1 = 1$  and  $\rho_n \geq 1/n$ . For any positive integer  $k \leq n$ , we have*

$$\mathbb{E}_{W_0} [\delta_\square(\tilde{f}_A, f'_0)] \leq 2\rho_n \inf_{W \in \mathcal{W}_2^+[k]} \delta_2(W, W'_0) + C \left[ \rho_n \|W_0\|_2 \sqrt{\frac{k}{n}} + \sqrt{\frac{\rho_n}{n}} \right], \quad (28)$$

$$\mathbb{E}_{W_0} [\delta_\square(\tilde{f}_{\tilde{\Theta}_\lambda}, f'_0)] \leq 2\rho_n \inf_{W \in \mathcal{W}_2^+[k]} \delta_2(W, W'_0) + C \left[ \rho_n \|W_0\|_2 \sqrt{\frac{k}{n}} + \sqrt{\frac{\rho_n}{n}} \right]. \quad (29)$$

If  $W_0$  belongs to some  $\mathcal{W}_2^+[k]$  or to  $\mathcal{W}_1^+[k, \mu]$  the convergence rates given by Proposition 7 are the same as the optimal rates for bounded graphons up to a  $\log^{-1/2}(k)$  factor. We conjecture that the  $\log^{-1/2}(k)$  factor should appear in Proposition 7. Indeed, for bounded graphons, this logarithmic terms derives from Szemerédi Regularity lemma and extensions of this lemma to  $L_p$  graphons have been recently proved [11]. Nevertheless, our arguments in the proof of Theorem 1 makes heavily

use of the boundedness of the graphons. In particular, one should replace all applications of McDiarmid's inequality (Lemma 1) by more involved concentration inequalities [13]. We leave this for future work.

When the graphons  $W_0$  is not a finite step graphons, a bias term is occurring in the risk bounds (26–29). As the estimation risk is measured in the cut-distance, one could have hoped to obtain a bias term in the cut distance also (instead of the larger  $l_1$  and  $l_2$  distances). It is an interesting open problem to prove whether one can obtain oracle inequalities with cut distance bias terms. Note that, for bounded graphons  $W \in \mathcal{W}^+$ , using Theorem 1, we can also get an oracle inequality with the  $\delta_1$  bias term and minimax optimal error term.

Upper bounds of the cut distance risk for  $L_p$  graphons estimation were previously obtained in [7] where the authors introduced the least cut norm estimator  $\hat{f}_{LC}$ . For any  $L_1$  normalized graphon  $W_0$  any  $\kappa \in [\log n/n, 1]$ , Borgs et al. [7] show in their Theorem 4.1 that this estimator  $\hat{f}_{LC}$  achieves the risk bound

$$\mathbb{E}_{W_0} [\delta_{\square}(\hat{f}_{LC}, f'_0)] \leq C \left[ \rho_n \inf_{W \in \mathcal{W}_1^+ [[\kappa]^{-1}, 1/2]} \delta_1(W, W'_0) + \rho_n \sqrt{\frac{\log n}{\kappa n}} + \sqrt{\frac{\rho_n}{n}} \right]. \quad (30)$$

For  $L_1$  graphons, this bound is quite similar (up to an additional  $\log^{1/2}(n)$  term) to those we obtained in (26–27) for the empirical estimators  $\tilde{f}_{\mathbf{A}}$  and  $\tilde{f}_{\tilde{\Theta}_\gamma}$ . Note that the least cut norm estimator can not be computed in polynomial time contrary to the empirical graphons associated to the adjacency matrix and to the singular value hard thresholding estimator. Also, when the true graphon  $W_0$  either belongs to  $\mathcal{W}_2^+$  or to  $\mathcal{W}^+[k]$ , then the rate in (30) is much slower than what has been obtained in Proposition 4 and Theorem 1.

## A Proof methods

In this section, we summarize some basic facts and fundamental results that we use in the proofs.

### A.1 Non-symmetric kernels

At some point, we will need to work with non-symmetric kernels and with kernel defined on general measurable subsets of  $\mathbb{R}$ . In this section we define the corresponding spaces. Let  $\mathcal{X}$  and  $\mathcal{Y}$  denote two bounded measurable subsets of  $\mathbb{R}$ . Then,  $\mathcal{W}_{\mathcal{X}, \mathcal{Y}}$  refers to the collection of bounded measurable functions  $W : \mathcal{X} \times \mathcal{Y} \rightarrow [-1, 1]$ . We will denote by  $\mathcal{W}_{\mathcal{X}, \mathcal{Y}}^+$  the collection of bounded measurable and non-negative functions  $W : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ . Let  $\mathcal{W}_{\mathcal{X}, \mathcal{Y}}[k]$  be the collection of  $k$ -step kernels, that is, the subset of kernels  $W \in \mathcal{W}_{\mathcal{X}, \mathcal{Y}}$  such that for some  $\mathbf{Q} \in \mathbb{R}^{k \times k}$  and some  $\phi_1 : \mathcal{X} \rightarrow [k]$ ,  $\phi_2 : \mathcal{Y} \rightarrow [k]$ ,

$$W(x, y) = \mathbf{Q}_{\phi_1(x), \phi_2(y)} \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}. \quad (31)$$

A kernel  $W$  is also said to be a  $q_1 \times q_2$ -step function when it decomposes as in (31) but where  $\mathbf{Q}$  is a size  $q_1 \times q_2$  matrix,  $\phi_1$  mapping  $\mathcal{X}$  to  $[q_1]$ , and  $\phi_2$  mapping  $\mathcal{Y}$  to  $[q_2]$ . The cut norm can be readily extended to kernels  $W \in \mathcal{W}_{\mathcal{X}, \mathcal{Y}}$  in the following way:

$$\|W\|_{\square} := \sup_{X \subset \mathcal{X}, Y \subset \mathcal{Y}} \left| \int_{X \times Y} W(x, y) dx dy \right| \quad (32)$$

where the supremum is taken over all measurable subsets  $X$  and  $Y$ .

## A.2 Concentration inequalities

In the proofs we repeatedly use Bernstein's inequality. We state it here for the readers' convenience. Let  $X_1, \dots, X_N$  be independent zero-mean random variables. Suppose that  $|X_i| \leq M$  almost surely, for all  $i$ . Then, for any  $t > 0$ ,

$$\mathbb{P} \left\{ \sum_{i=1}^N X_i \geq t \right\} \leq \exp \left[ - \frac{t^2}{2 \sum_i \mathbb{E}[X_i^2] + 2Mt/3} \right]. \quad (33)$$

We shall also rely on the bounded difference inequality (also called Mc Diarmid inequality).

**Lemma 1** (Bounded difference inequality). *Let  $X_1, \dots, X_n$  denote  $n$  independent real random variables. Assume that  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is a measurable function satisfying, for some positive constants  $(c_i)_{1 \leq i \leq n}$ , the bounded difference condition*

$$|g(x_1, \dots, x_i, \dots, x_n) - g(x_1, \dots, x'_i, \dots, x_n)| \leq c_i,$$

*for all  $x = (x_1, \dots, x_i, \dots, x_n) \in \mathbb{R}^n$ ,  $x' = (x_1, \dots, x'_i, \dots, x_n) \in \mathbb{R}^n$  and all  $i \in [n]$ . Then, the random variable  $Z = g(X_1, \dots, X_n)$  satisfies*

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq \exp \left[ - \frac{2t^2}{\sum_{i=1}^n c_i^2} \right],$$

*for all  $t > 0$ .*

## A.3 Fano's lemma

In the sequel,  $\mathcal{KL}(\cdot, \cdot)$  denotes the Kullback-Leibler divergence between two distributions. In this manuscript, all the proofs of the minimax lower bounds rely on Fano's method. The following version of Fano's lemma is borrowed from [32]:

**Lemma 2.** [32, Theorem 2.7] *Consider a parametric model  $\mathbb{P}_\theta$ , with  $\theta \in \Theta$  and a metric  $d(\cdot, \cdot)$  on  $\Theta$ . Assume that  $\Theta$  contains elements  $\theta_1, \dots, \theta_M$ ,  $M \geq 3$ , such that for all  $j, k \in [M]$  with  $j \neq k$*

$$(i) \quad d(\theta_j, \theta_k) \geq s > 0,$$

$$(ii) \quad \mathcal{KL}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_k}) \leq \log(M)/32.$$

*Then, we have*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} [d(\hat{\theta}, \theta)] \geq Cs,$$

*where the constant  $C > 0$  is numeric.*

## A.4 Khintchine's inequality

Next, we state a particular case of Khintchine's inequality that turns out to be useful for bounding the cut norm of step kernels in terms of their  $l_1$  norm:

**Lemma 3.** [30] *Let  $\epsilon_1, \dots, \epsilon_p$  be i.i.d. Rademacher random variables and let  $x_1, \dots, x_p$  be some real numbers. Then,*

$$\mathbb{E} \left[ \left| \sum_{i=1}^p \epsilon_i x_i \right| \right] \geq \frac{1}{\sqrt{2}} \left[ \sum_{i=1}^p x_i^2 \right]^{1/2}. \quad (34)$$



We use this result to prove the following lower bound on the cut norm of step kernels:

**Lemma 4.** *Let  $U : \mathcal{X} \times \mathcal{Y} \mapsto [-1, 1]$  denote a measurable  $q_1 \times q_2$ -step function. Then,*

$$\|U\|_{\square} \geq \frac{1}{4\sqrt{2q_2}} \|U\|_1. \quad (35)$$

*Proof of Lemma 4.* There exist partitions  $\mathcal{X} = \mathcal{X}_1 \cup \dots \mathcal{X}_{q_1}$  and  $\mathcal{Y} = \mathcal{Y}_1 \cup \dots \mathcal{Y}_{q_2}$  such that, for any fixed  $y \in \mathcal{Y}$ ,  $U(x, y)$  is constant over  $\mathcal{X}_i$  for all  $i \in [q_1]$  and, for any fixed  $x \in \mathcal{X}$ ,  $U(x, y)$  is constant over  $\mathcal{Y}_i$  for all  $i \in [q_2]$ . For any  $a \in [q_1]$  (resp.  $b \in [q_2]$ ), denote  $x_a$  (resp.  $y_b$ ) any element of  $\mathcal{X}_a$  (resp.  $\mathcal{Y}_b$ ). By definition of  $\|U\|_{\square}$ ,

$$\begin{aligned} \|U\|_{\square} &= \sup_{S \subset \mathcal{X}, T \subset \mathcal{Y}} \left| \int_{S, T} U(x, y) dx dy \right| \\ &= \sup_{S \subset \mathcal{X}, T \subset \mathcal{Y}} \sum_{a=1}^{q_1} \sum_{b=1}^{q_2} \left| \lambda(S \cap \mathcal{X}_a) \lambda(T \cap \mathcal{Y}_b) U(x_a, y_b) \right| \\ &= \sup_{\epsilon \in \{0,1\}^{q_1}} \sup_{\epsilon' \in \{0,1\}^{q_2}} \left| \sum_{a=1}^{q_1} \sum_{b=1}^{q_2} \epsilon_a \lambda(\mathcal{X}_a) \epsilon'_b \lambda(\mathcal{Y}_b) U(x_a, y_b) \right|, \end{aligned}$$

where we used in the last line that the value of the sum only depends on  $S$  and  $T$  through the quantities  $\lambda(S \cap \mathcal{X}_a)$  and  $\lambda(T \cap \mathcal{Y}_b)$ . Since the maximum of a linear function on a convex set is achieved at an extremal point, it follows that

$$\begin{aligned} \|U\|_{\square} &= \sup_{\epsilon \in \{0,1\}^{q_1}, \epsilon' \in \{0,1\}^{q_2}} \left| \sum_{a=1}^{q_1} \sum_{b=1}^{q_2} \epsilon_a \lambda(\mathcal{X}_a) \epsilon'_b \lambda(\mathcal{Y}_b) U(x_a, y_b) \right| \\ &\geq \frac{1}{4} \sup_{\epsilon \in \{-1,1\}^{q_1}, \epsilon' \in \{-1,1\}^{q_2}} \left| \sum_{a \in [q_1], b \in [q_2]} \epsilon_a \epsilon'_b \lambda(\mathcal{X}_a) \lambda(\mathcal{Y}_b) U[x_a, y_b] \right| \\ &\geq \frac{1}{4} \sup_{\epsilon' \in \{-1,1\}^{q_2}} \sum_{a \in [q_1]} \lambda(\mathcal{X}_a) \left| \sum_{b \in [q_2]} \epsilon'_b \lambda(\mathcal{Y}_b) U[x_a, y_b] \right| \end{aligned}$$

where we use (8) and take  $\epsilon_a = \text{sign} \sum_{b \in [q_2]} \epsilon'_b \lambda(\mathcal{Y}_b) U[x_a, y_b]$ . Let  $v = (v_1, \dots, v_{q_2})$  denote i.i.d. Rademacher random variables and let  $\mathbb{E}_v[\cdot]$  denotes the expectation with respect to  $v$ . Now, Khintchine's inequality (34) and Cauchy-Schwarz inequality imply

$$\begin{aligned} \sup_{\epsilon' \in \{-1,1\}^{q_2}} \sum_{a \in [q_1]} \lambda(\mathcal{X}_a) \left| \sum_{b \in [q_2]} \epsilon'_b \lambda(\mathcal{Y}_b) U[x_a, y_b] \right| &\geq \mathbb{E}_v \left[ \sum_{a \in [q_1]} \lambda(\mathcal{X}_a) \left| \sum_{b \in [q_2]} v_b \lambda(\mathcal{Y}_b) U[x_a, y_b] \right| \right] \\ &\geq \frac{1}{\sqrt{2}} \sum_{a \in [q_1]} \lambda(\mathcal{X}_a) \left( \sum_{b \in [q_2]} \lambda^2(\mathcal{Y}_b) U^2[x_a, y_b] \right)^{1/2} \\ &\geq \frac{1}{\sqrt{2q_2}} \sum_{a \in [q_1]} \sum_{b \in [q_2]} \lambda(\mathcal{X}_a) \lambda(\mathcal{Y}_b) |U[x_a, y_b]| \\ &= \frac{1}{\sqrt{2q_2}} \|U\|_1. \end{aligned}$$

□

## B Proof of Proposition 2

Since the diagonals of  $\mathbf{A}$  and  $\mathbf{\Theta}$  are both zero, it suffices to control the supremum over disjoint subsets  $S$  and  $T$  (see, e.g., [8])

$$\|\mathbf{A} - \mathbf{\Theta}_0\|_{\square} \leq \frac{4}{n^2} \max_{S \cap T = \emptyset} \left| \sum_{i \in S, j \in T} (\mathbf{A}_{ij} - \mathbf{\Theta}_{ij}) \right|.$$

Let  $S$  and  $T$  be any two disjoint subsets of  $[n]$ . Using Bernstein's inequality (33) we have that

$$\begin{aligned} \mathbb{P} \left\{ \left| \sum_{i \in S, j \in T} \mathbf{A}_{ij} - \mathbf{\Theta}_{ij} \right| \geq 3\sqrt{(\|\mathbf{\Theta}_0\|_1 + n)n} \right\} &\leq 2 \exp \left( -\frac{9(\|\mathbf{\Theta}_0\|_1 + n)n}{2\|\mathbf{\Theta}_0\|_1 + 2\sqrt{(\|\mathbf{\Theta}_0\|_1 + n)n}} \right) \\ &\leq 2 \exp \left( -\frac{9}{4}n \right) \end{aligned}$$

Now, using that the number of disjoint pairs  $(S, T)$  is  $3^n$  and the union bound, we get that the probability that  $|\sum_{i \in S, j \in T} \mathbf{A}_{ij} - \mathbf{\Theta}_{ij}|$  exceeds  $3\sqrt{(\|\mathbf{\Theta}_0\|_1 + n)n}$  for some  $(S, T)$  is bounded by  $2 \exp(-n)$ . Hence, we have

$$\|\mathbf{A} - \mathbf{\Theta}_0\|_{\square} \leq 4 \sup_{S \cap T = \emptyset} \frac{1}{n^2} \left| \sum_{(i,j) \in S \times T} \mathbf{A}_{ij} - \mathbf{\Theta}_{ij} \right| \leq 12 \sqrt{\frac{\|\mathbf{\Theta}_0\|_1 + n}{n^3}}$$

with probability  $1 - 2e^{-n}$ . Now bounding the distance by 1 in the exceptional case we get the statement of Proposition 2.

## C Proof of Proposition 3

This proof is based on Fano's method. To apply Fano's Lemma (Lemma 2), it is enough to check that there exists a finite subset  $\Omega$  of  $\mathcal{T}[2, \rho_n]$  such that for any two distinct  $\mathbf{\Theta}, \mathbf{\Theta}'$  in  $\Omega$  we have

$$(a) \quad \|\mathbf{\Theta} - \mathbf{\Theta}'\|_{\square} \geq C \sqrt{\rho_n} \left( \frac{1}{\sqrt{n}} \wedge \sqrt{\rho_n} \right) \text{ and}$$

$$(b) \quad \mathcal{KL}(\mathbb{P}_{\mathbf{\Theta}}, \mathbb{P}_{\mathbf{\Theta}'}) \leq \log(|\Omega|)/32$$

for some constants  $C > 0$ .

To prove it, we fix some  $\rho_n/4 > \epsilon > 0$ . For any  $u \in \{-1, 1\}^n$ , define  $\mathbf{\Theta}_u$  by  $(\mathbf{\Theta}_u)_{i,j} = \rho_n/2 + u(i)u(j)\epsilon$  where  $u = (u(1), \dots, u(n))$ . Obviously, we have

$$\{\mathbf{\Theta}_u : u \in \{-1, 1\}^n\} \subset \mathcal{T}[2, \rho_n].$$

Denote  $V_u := \{i \in [n] : u(i) = 1\}$  and  $\bar{V}_u$  its complementary. Then, if we take  $S := V_u \setminus V_v$  and  $T := V_v \cap V_u$ , we obtain

$$\left| \sum_{i \in S, j \in T} (\mathbf{\Theta}_u - \mathbf{\Theta}_v)_{ij} \right| = 2\epsilon |V_u \setminus V_v| |V_v \cap V_u|.$$

By symmetry, we derive that

$$\begin{aligned} n^2 \|\mathbf{\Theta}_u - \mathbf{\Theta}_v\|_{\square} &\geq 2\epsilon \max\{|V_u \setminus V_v|, |V_v \setminus V_u|\} \max\{|\bar{V}_u \cap \bar{V}_v|, |V_v \cap V_u|\} \\ &\geq \frac{\epsilon}{2} |V_u \Delta V_v| (n - |V_u \Delta V_v|), \end{aligned}$$

where  $A \triangle B$  is the symmetric difference of  $A$  and  $B$ . We can use Varshamov-Gilbert bound (see, e.g., [32, Lemma 2.9]) to pick  $u_1, \dots, u_N$  satisfying

$$\frac{n}{4} \leq |V_{u_i} \triangle V_{u_j}| \leq \frac{3n}{4} \quad \text{for } i \neq j \in [N]$$

with  $N \geq \exp(c_1 n)$  for some  $c_1 > 0$ . Let  $\Omega = \{\Theta_{u_i} : i = 1, \dots, N\}$ , hence we have  $\log |\Omega| \geq c_1 n$  and

$$\|\Theta_{u_i} - \Theta_{u_j}\|_{\square} \geq \epsilon/14$$

which proves (a) when one takes  $\epsilon$  as defined in (36) below.

To prove (b) we use the definition of Kullback-Leibler divergence  $\mathcal{KL}(\mathbb{P}_{\Theta_u}, \mathbb{P}_{\Theta_v})$  and  $\log x \leq x-1$  for  $x > 0$  to get

$$\begin{aligned} \mathcal{KL}(\mathbb{P}_{\Theta_u}, \mathbb{P}_{\Theta_v}) &= \sum_{ij} (\Theta_u)_{i,j} \log \left( \frac{(\Theta_u)_{i,j}}{(\Theta_v)_{i,j}} \right) + (1 - (\Theta_u)_{i,j}) \log \left( \frac{(1 - \Theta_u)_{i,j}}{1 - (\Theta_v)_{i,j}} \right) \\ &\leq \sum_{ij} \frac{((\Theta_u)_{i,j} - (\Theta_v)_{i,j})^2}{(\Theta_v)_{i,j} (1 - (\Theta_v)_{i,j})}. \end{aligned}$$

Now,  $(\Theta_v)_{i,j} \geq \rho_n/4$  and  $\rho_n \leq 1$  imply

$$\mathcal{KL}(\mathbb{P}_{\Theta_{u_i}}, \mathbb{P}_{\Theta_{u_j}}) \leq \frac{16}{3\rho_n} \sum_{ij} ((\Theta_u)_{i,j} - (\Theta_v)_{i,j})^2 \leq \frac{16n^2\epsilon^2}{3\rho_n}.$$

Taking

$$\epsilon = c_2 \sqrt{\rho_n} \left( \frac{1}{\sqrt{n}} \wedge \sqrt{\rho_n} \right) \quad (36)$$

with a suitable constant  $c_2 > 0$ , we have that

$$\mathcal{KL}(\mathbb{P}_{\Theta_{u_i}}, \mathbb{P}_{\Theta_{u_j}}) \leq \log |\Omega|/32$$

which proves (b).

## D Proof of Proposition 4

Set  $\mathbf{E} = \mathbf{A} - \Theta_0$ . We have the following simple proposition (see Theorem 5 in [25])

**Proposition 8.** *If  $\lambda \geq \|\mathbf{E}\|_{2 \rightarrow 2}$ , then*

$$\|\tilde{\Theta}_\lambda - \Theta_0\|_{2 \rightarrow 2} \leq 2\lambda.$$

In view of Proposition 8 we need to estimate  $\|\mathbf{E}\|$  with high probability in order to specify the value of the regularization parameter  $\lambda$ . Let  $\mathbf{E}^* = (\mathbf{E}_{ij}^*)$  be such that  $\mathbf{E}_{ij}^* = \mathbf{E}_{ij}$  for  $i < j$  and  $\mathbf{E}_{ij}^* = 0$  for  $i \geq j$ . Then  $\|\mathbf{E}\|_{2 \rightarrow 2} \leq 2\|\mathbf{E}^*\|$ . We can upper bound  $\|\mathbf{E}^*\|$  using the following bound on the spectral norm of random matrices from [3]:

**Proposition 9.** *Let  $\mathbf{W}$  be the  $n \times m$  rectangular matrix whose entries  $\mathbf{W}_{ij}$  are independent centered random variables bounded (in absolute value) by some  $\sigma_* > 0$ . Then, for any  $0 < \epsilon \leq 1/2$  there exists a universal constant  $c_\epsilon$  such that, for every  $t \geq 0$*

$$\mathbb{P} \left\{ \|\mathbf{W}\|_{2 \rightarrow 2} \geq (1 + \epsilon) 2\sqrt{2}(\sigma_1 \vee \sigma_2) + t \right\} \leq (n \wedge m) \exp \left( \frac{-t^2}{c_\epsilon \sigma_*^2} \right)$$

where we have defined

$$\sigma_1 = \max_i \sqrt{\sum_j \mathbb{E}[\mathbf{W}_{ij}^2]}, \quad \sigma_2 = \max_j \sqrt{\sum_i \mathbb{E}[\mathbf{W}_{ij}^2]}.$$

For  $\mathbf{E}^*$ , we have  $\sigma_1 \leq \sqrt{\rho_n n}$ ,  $\sigma_2 \leq \sqrt{\rho_n n}$ , and  $\sigma_* \leq 1$ . Taking  $\epsilon = 1/2$  and  $t = \sqrt{2c_\epsilon \log(n)}$  in Proposition 9, we obtain that there exists absolute constants  $c^*$  such that

$$\|\mathbf{E}\|_{2 \rightarrow 2} \leq 2 \|\mathbf{E}^*\|_{2 \rightarrow 2} \leq 6\sqrt{2\rho_n n} + 2c^* \sqrt{\log(n)}, \quad (37)$$

with probability at least  $1 - 1/n$ . Since  $\rho_n \geq \log(n)/n$ , we can take  $\lambda = c\sqrt{\rho_n n}$  where  $c \geq 12\sqrt{2} + 4c^*$  so that  $\|\mathbf{E}\|_{2 \rightarrow 2} \leq \lambda/2$ . Then, Proposition 8 implies

$$\|\tilde{\Theta}_\lambda - \Theta_0\|_{2 \rightarrow 2} \leq C\sqrt{\rho_n n}.$$

It is easy to see that the cut-norm of a matrix can be bounded by its spectral norm:

$$\|\mathbf{A}\|_\square \leq \frac{1}{n} \|\mathbf{A}\|_{2 \rightarrow 2}.$$

Bound on the cut-norm (15) then follows from

$$\|\tilde{\Theta}_\lambda - \Theta_0\|_\square \leq \frac{1}{n} \|\tilde{\Theta}_\lambda - \Theta_0\|_{2 \rightarrow 2} \leq C\sqrt{\frac{\rho_n}{n}}.$$

In order to prove the Frobenius bound (14), we use the argument from [25]: we can equivalently write the singular value hard thresholding estimator as the solution to the following optimization problem:

$$\tilde{\Theta}_\lambda \in \arg \min_{\Theta \in \mathbb{R}^{n \times n}} \{ \|\mathbf{A} - \Theta\|_2^2 + \lambda^2 \text{rank}(\Theta) \}$$

which implies that, with probability larger than  $1 - 1/n$ ,

$$\begin{aligned} \|\tilde{\Theta}_\lambda - \Theta_0\|_2^2 &\leq 2|\langle \mathbf{E}, \tilde{\Theta}_\lambda - \Theta_0 \rangle| + \lambda^2 \text{rank}(\Theta_0) - \lambda^2 \text{rank}(\tilde{\Theta}_\lambda) \\ &\leq 2 \|\mathbf{E}\|_{2 \rightarrow 2} \left\| \tilde{\Theta}_\lambda - \Theta_0 \right\|_2 \sqrt{\text{rank}(\tilde{\Theta}_\lambda - \Theta_0) + \lambda^2 \text{rank}(\Theta_0) - \lambda^2 \text{rank}(\tilde{\Theta}_\lambda)} \\ &\leq \frac{1}{2} \|\tilde{\Theta}_\lambda - \Theta_0\|_2^2 + 2\lambda^2 \text{rank}(\Theta_0), \end{aligned}$$

where we used in the last line that  $\|\mathbf{E}\|_{2 \rightarrow 2} \leq \lambda/2$ . Since  $\text{rank}(\Theta_0) \leq k$ , we have proved (14).

## E Proof of Theorem 1

Note that both  $f_0 = \rho_n W_0$  and  $\tilde{f}_{\Theta_0}$  are proportional to  $\rho_n$ , so without loss of generality we can assume that  $\rho_n = 1$ . For  $k \geq n/2$ , the result is a straightforward consequence of the second Sampling Lemma for Graphons of [28] stated in Proposition 1. Given any graphon  $W_0 \in \mathcal{W}^+[k]$ , one can always divide some of the steps into smaller in such a way that  $W_0$  is a  $2k$ -step graphon whose weights are all less than or equal to  $1/k$ . Thus, we only need to prove the results for all graphons  $W_0 \in \mathcal{W}^+[k]$  with  $32 \leq k \leq n$  and such that its weights are all smaller or equal to  $2/k$ .

Let  $\Theta'_0$  be the matrix with entries  $(\Theta'_0)_{ij} = W(\xi_i, \xi_j)$  for all  $i, j$ . As opposed to  $\Theta_0$ , the diagonal entries of  $\Theta'_0$  are not constrained to be null. By the triangle inequality, we have

$$\mathbb{E} \left[ \delta_\square \left( \tilde{f}_{\Theta_0}, W_0 \right) \right] \leq \mathbb{E} \left[ \delta_\square \left( \tilde{f}_{\Theta_0}, \tilde{f}_{\Theta'_0} \right) \right] + \mathbb{E} \left[ \delta_\square \left( \tilde{f}_{\Theta'_0}, W_0 \right) \right]. \quad (38)$$

As the entries of  $\Theta_0$  coincide with those of  $\Theta'_0$  outside the diagonal, the difference  $\tilde{f}_{\Theta_0} - \tilde{f}_{\Theta'_0}$  is null outside of a set of measure  $1/n$ . Since  $\|W_0\|_\infty \leq 1$ ,  $\mathbb{E}[\delta_\square(\tilde{f}_{\Theta_0}, \tilde{f}_{\Theta'_0})] \leq 1/n$ . Thus, we only need to prove that

$$\mathbb{E}[\delta_\square(\tilde{f}_{\Theta'_0}, W_0)] \leq C \sqrt{\frac{k}{n \log(k)}}. \quad (39)$$

We first need to build two suitable representations of  $W_0$  and  $\tilde{f}_{\Theta'_0}$  in the quotient space  $\widetilde{W}^+$ .

**Step 1:** *Construction of a suitable representation  $W$  of  $W_0$  in  $\widetilde{W}^+$ .*

In the sequel, we denote  $q_1 := \lfloor \sqrt{k} \rfloor$ . Here, we want to choose  $W$  in such a way that a distortion of  $W$  is well approximated in the cut norm by a  $q_1$ -step kernel. We use the following lemma which is based on a variation of Szemerédi lemma. Let  $\mathbf{Q}_0 \in \mathbb{R}_{\text{sym}}^{k \times k}$  and  $\phi_0 : [0, 1] \rightarrow [k]$  be associated to  $W_0$  as in definition (17).

**Lemma 5.** *There exist a permutation  $\pi$  of  $[k]$  and a partition  $\mathcal{P} = (P_1, \dots, P_{q_1})$  of  $[k]$  made of successive intervals such that the following holds. Let  $\mathbf{Q}$  be the matrix obtained from  $\mathbf{Q}_0$  by jointly applying the permutation  $\pi$  to its rows and its columns. Denote by  $\phi = \pi \circ \phi_0$ , and for  $a = 1, \dots, k$ ,  $\lambda_a := \lambda(\phi^{-1}(a))$ . There are two matrices  $\mathbf{Q}^{(ap)}$  and  $\mathbf{Q}^{(ap,+)} \in [0, 1]^{k \times k}$  that are  $q_1$ -block-constant according to the partition  $\mathcal{P}$  and that satisfy*

$$\sup_{\epsilon \in \{0,1\}^k, \epsilon' \in \{0,1\}^k} \left| \sum_{a,b=1}^k \epsilon_a \epsilon'_b \lambda_b \sqrt{\lambda_a} (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)}) \right| \leq C \sqrt{\frac{k}{\log(k)}}, \quad (40)$$

$$\sup_{\epsilon \in \{0,1\}^k, \epsilon' \in \{0,1\}^k} \left| \sum_{a,b=1}^k \epsilon_a \epsilon'_b \sqrt{\lambda_b} \sqrt{\lambda_a} (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap,+)}) \right| \leq C \frac{k}{\sqrt{\log(k)}}. \quad (41)$$

Invoking Lemma 5, we consider the graphons

$$W(x, y) := \mathbf{Q}_{\phi(x)\phi(y)}, \quad W_1(x, y) := \mathbf{Q}_{\phi(x)\phi(y)}^{(ap)}, \quad W_1^+(x, y) := \mathbf{Q}_{\phi(x)\phi(y)}^{(ap,+)}. \quad (42)$$

Obviously,  $W$  is weakly isomorphic to  $W_0$ .

**Step 2:** *Construction of a suitable representation  $\widehat{W}$  of  $\tilde{f}_{\Theta'_0}$  in the quotient space  $\widetilde{W}^+$ .*

Recall that  $\xi_1, \dots, \xi_n$  are the i.i.d. uniformly distributed random variables in the  $W$ -random graph model (1) and that  $\phi$  is defined in the previous step. For  $a = 1, \dots, k$ , let

$$\widehat{\lambda}_a = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\xi_i \in \phi^{-1}(a)\}}$$

be the (unobserved) empirical frequency of the group  $a$  corresponding to a finer partition of  $[0, 1]$  given by  $\phi$ . For  $a = 1, \dots, q_1$ , let

$$\widehat{\omega}_a = \frac{1}{n} \sum_{i=1}^n \sum_{b \in P_a} \mathbb{1}_{\{\xi_i \in \phi^{-1}(b)\}}$$

be the (unobserved) empirical frequency of the group  $a$  corresponding to a coarser partition  $P$  of  $[0, 1]$  given by  $\mathcal{P} \circ \phi$ .

The relations  $\sum_{a=1}^k \lambda_a = \sum_{a=1}^k \widehat{\lambda}_a = 1$  imply

$$\sum_{a: \lambda_a > \widehat{\lambda}_a} (\lambda_a - \widehat{\lambda}_a) = \sum_{a: \widehat{\lambda}_a > \lambda_a} (\widehat{\lambda}_a - \lambda_a) \quad \text{and} \quad \sum_{a: \omega_a > \widehat{\omega}_a} (\omega_a - \widehat{\omega}_a) = \sum_{a: \widehat{\omega}_a > \omega_a} (\widehat{\omega}_a - \omega_a). \quad (43)$$

Consider a function  $\psi : [0, 1] \rightarrow [k]$  such that:

- (i) For all  $a \in [k]$ ,  $\lambda(\{x, \psi(x) = \phi(x) = a\}) = \widehat{\lambda}_a \wedge \lambda_a$ ,
- (ii) for all  $a \in [q_1]$ ,  $\lambda\left[\{x, \psi(x) \in P_a \text{ and } \phi(x) \in P_a\}\right] = \omega_a \wedge \widehat{\omega}_a$ ,
- (iii) for all  $a \in [k]$ ,  $\lambda(\psi^{-1}(a)) = \widehat{\lambda}_a$ .

Such a function  $\psi$  exists. First we construct  $\psi$  to satisfy (i) and (iii). For each  $a$  such that  $\lambda_a > \widehat{\lambda}_a$ , conditions (i) and (iii) are trivially satisfied if we take  $\psi^{-1}(a)$  to be any subset of  $\phi^{-1}(a)$  of Lebesgue measure  $\widehat{\lambda}_a$  and there is a subset of  $\phi^{-1}(a)$  of Lebesgue measures  $\lambda_a - \widehat{\lambda}_a$  left non-assigned. Summing over all such  $a$ , we see that there is a union of subsets with Lebesgue measure  $m_+ := \sum_{a: \lambda_a > \widehat{\lambda}_a} (\lambda_a - \widehat{\lambda}_a)$  left non-assigned. On the other hand, for  $a$  such that  $\lambda_a < \widehat{\lambda}_a$ , we must have  $\psi(x) = a$  for  $x \in \phi^{-1}(a)$  to satisfy (i), while to meet condition (iii) we need additionally to assign  $\psi(x) = a$  for  $x$  on a set of Lebesgue measure  $\widehat{\lambda}_a - \lambda_a$ . Summing over all such  $a$ , we need additionally to find a set of Lebesgue measure  $m_- := \sum_{a: \widehat{\lambda}_a > \lambda_a} (\widehat{\lambda}_a - \lambda_a)$  to make such assignments. But this set is readily available as the union of non-assigned intervals for all  $a$  such that  $\lambda_a > \widehat{\lambda}_a$  since  $m_+ = m_-$  by virtue of (43). To ensure that condition (ii) is satisfied, we assign as a priority  $\psi(x)$  to values belonging to the same partition element as  $\phi(x)$ . Again, (43) ensures that this is possible.

Finally, define the graphons  $\widehat{W}(x, y) = \mathbf{Q}_{\psi(x), \psi(y)}$ ,  $\widehat{W}_1(x, y) = \mathbf{Q}_{\psi(x), \psi(y)}^{(ap)}$ , and  $\widehat{W}_1^+(x, y) = \mathbf{Q}_{\psi(x), \psi(y)}^{(ap, +)}$  where  $\mathbf{Q}$ ,  $\mathbf{Q}^{(ap)}$ , and  $\mathbf{Q}^{(ap, +)}$  are as in (42). Notice that in view of (iii)  $\widehat{W}$  is weakly isomorphic to the empirical graphon  $\widetilde{f}_{\Theta'_0}$ . Let  $\mathcal{R} = \{x, \phi(x) \neq \psi(x)\}$ . Since  $W$  and  $\widehat{W}$  match on  $\mathcal{R}^c \times \mathcal{R}^c$ , the purpose of (i) is to minimize the Lebesgue measure of the support of  $W - \widehat{W}$ . With properties (i) and (iii) alone, it would be possible to prove that  $\mathbb{E}[\|W - \widehat{W}\|_{\square}] \leq C\sqrt{k/n}$  as the Lebesgue measure of its support is at most of order  $\sqrt{k/n}$ . We will improve this rate by a logarithmic term as (ii) will enforce that the cut norm of  $W - \widehat{W}$  is much smaller than its Lebesgue measure.

**Step 3: Control of the cut norm.** Since  $\delta_{\square}(\cdot, \cdot)$  is a metric on the quotient space  $\widetilde{\mathcal{W}}^+$ ,

$$\delta_{\square}(W_0, \widetilde{f}_{\Theta'_0}) \leq \|W - \widehat{W}\|_{\square} = \sup_{S, T} \left| \int_{S \times T} (W(x, y) - \widehat{W}(x, y)) dx dy \right|.$$

By definition of  $\psi$ , the two functions  $W(x, y)$  and  $\widehat{W}(x, y)$  are equal except possibly when either  $x$  or  $y$  belongs to  $\mathcal{R}$ . As a consequence of triangular inequality and of the symmetry of  $W - \widehat{W}$ , we get

$$\begin{aligned} \|W - \widehat{W}\|_{\square} &\leq 2 \sup_{S \subset \mathcal{R}, T \subset \mathcal{R}^c} \left| \int_{S \times T} (W(x, y) - \widehat{W}(x, y)) dx dy \right| \\ &\quad + \sup_{S, T \subset \mathcal{R}} \left| \int_{S \times T} (W(x, y) - \widehat{W}(x, y)) dx dy \right| \\ &= 2 \left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}^c} \right\|_{\square} + \left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}} \right\|_{\square}. \end{aligned} \quad (44)$$

First, we focus on  $\mathbb{E}[\|(W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}^c}\|_{\square}]$ , the second term being handled similarly at the end of the proof. For  $a$  and  $b$  in  $[k]$ , we write  $a \sim_P b$  (resp.  $a \asymp_P b$ ) when  $a$  and  $b$  belongs (resp. do not belong) to the same element of the partition  $P$ . Define

$$\mathcal{R}_2 := \{x, \psi(x) \asymp_P \phi(x)\}.$$



Obviously, we have  $\mathcal{R}_2 \subset \mathcal{R}$ . Property (ii) of  $\psi$ , implies that  $\lambda(\mathcal{R}_2) = \sum_{a=1}^{q_1} (\omega_a - \widehat{\omega}_a)_+$ . We shall rely on the decomposition  $W = W_1 + (W - W_1)$  and  $\widehat{W} = \widehat{W}_1 + (\widehat{W} - \widehat{W}_1)$ . For any  $x \in \mathcal{R} \setminus \mathcal{R}_2$ , we have by definition (42) of  $W_1$  that  $(W_1 - \widehat{W}_1)(x, y) = 0$ . Together with the triangular inequality, this yields

$$\left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}^c} \right\|_{\square} \leq \left\| (W_1 - \widehat{W}_1)|_{\mathcal{R}_2 \times \mathcal{R}^c} \right\|_{\square} + \left\| (W - W_1)|_{\mathcal{R} \times \mathcal{R}^c} \right\|_{\square} + \left\| (\widehat{W} - \widehat{W}_1)|_{\mathcal{R} \times \mathcal{R}^c} \right\|_{\square}. \quad (45)$$

To control the first expression in the rhs, we simply bound the cut norm of the difference by its  $l_1$  norm

$$\left\| (W_1 - \widehat{W}_1)|_{\mathcal{R}_2 \times \mathcal{R}^c} \right\|_{\square} \leq \left\| (W_1 - \widehat{W}_1)|_{\mathcal{R}_2 \times \mathcal{R}^c} \right\|_1 \leq \lambda(\mathcal{R}_2) \left\| W_1 - \widehat{W}_1 \right\|_{\infty} \leq \lambda(\mathcal{R}_2),$$

since  $W_1$  and  $\widehat{W}_1$  take values in  $[0, 1]$ . Then, relying on the fact that  $n\widehat{\omega}_a$  is distributed as a Binomial random variable with parameters  $(n, \omega_a)$  and on Cauchy-Schwarz inequality, we get  $\mathbb{E} |\omega_a - \widehat{\omega}_a| \leq \sqrt{\frac{\omega_a(1-\omega_a)}{n}}$  and

$$\begin{aligned} \mathbb{E} \left[ \left\| (W_1 - \widehat{W}_1)|_{\mathcal{R}_2 \times \mathcal{R}^c} \right\|_{\square} \right] &\leq \mathbb{E} \left[ \sum_{a=1}^{q_1} |\omega_a - \widehat{\omega}_a| \right] \\ &\leq \sum_{a=1}^{q_1} \sqrt{\frac{\omega_a(1-\omega_a)}{n}} \leq \sqrt{\frac{q_1}{n}} \leq \frac{k^{1/4}}{\sqrt{n}}, \end{aligned} \quad (46)$$

where we used again Cauchy-Schwarz in the last line. Let us turn to the second and third expressions in (45). To this end, we introduce a new kernel function  $U$ . For  $a = 1, \dots, k$ , define  $\widehat{\lambda}_a^\delta = |\lambda_a - \widehat{\lambda}_a|$  and the functions  $F_{\widehat{\lambda}^\delta} : [k] \rightarrow [0, \sum_a |\lambda_a - \widehat{\lambda}_a|]$  and  $F_\phi : [k] \mapsto [0, 1]$  by

$$\begin{aligned} F_\phi(b) &= \sum_{a=1}^b \lambda_a \quad \text{and set} \quad F_\phi(0) = 0 \\ F_{\widehat{\lambda}^\delta}(b) &= \sum_{a=1}^b \widehat{\lambda}_a^\delta \quad \text{and set} \quad F_{\widehat{\lambda}^\delta}(0) = 0. \end{aligned} \quad (47)$$

For any  $a, b \in [k]$ , set  $\widehat{\Pi}_{a,b} = [F_{\widehat{\lambda}^\delta}(a-1), F_{\widehat{\lambda}^\delta}(a)) \times [F_\phi(b-1), F_\phi(b))$  and let  $U$  be a  $k \times k$  step kernel on  $[0, \sum_a |\widehat{\lambda}_a - \lambda_a|] \times [0, 1]$  defined by

$$U(x, y) := \sum_{a,b=1}^k \left[ \mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)} \right] \mathbb{1}_{\widehat{\Pi}_{a,b}}(x, y).$$

By definition of  $\mathcal{R}$  and of the function  $\psi$ , we have that for any  $a \in [k]$ ,  $\lambda(\phi^{-1}(a)) \cap \mathcal{R} = (\lambda_a - \widehat{\lambda})_+$  and  $\lambda(\psi^{-1}(a)) \cap \mathcal{R}^c = \lambda_a \wedge \widehat{\lambda}$ . As a consequence, the restriction of  $(W - W_1)$  to  $\mathcal{R} \times \mathcal{R}^c$  is, up to a measure preserving bijection of its rows and of its columns, equal to the restriction of  $U$  to the set  $(\cup_{a: \lambda_a > \widehat{\lambda}_a} [F_{\widehat{\lambda}^\delta}(a-1), F_{\widehat{\lambda}^\delta}(a))) \times (\cup_a [F_\phi(a-1), F_\phi(a-1) + \widehat{\lambda}_a \wedge \lambda_a])$ . This entails that

$$\left\| (W - W_1)|_{\mathcal{R} \times \mathcal{R}^c} \right\|_{\square} \leq \|U\|_{\square}. \quad (48)$$

On the other hand, for any  $(x, y) \in \mathcal{R} \times \mathcal{R}^c$ ,

$$(\widehat{W} - \widehat{W}_1)(x, y) = \mathbf{Q}_{\psi(x)\psi(y)} - \mathbf{Q}_{\psi(x)\psi(y)}^{(ap)} = \mathbf{Q}_{\psi(x)\phi(y)} - \mathbf{Q}_{\psi(x)\phi(y)}^{(ap)}$$

by the definition of  $\mathcal{R}$ . In view of the definition of  $\psi$ , for any  $a \in [k]$  we have  $\lambda(\phi^{-1}(a)) \cap \mathcal{R} = (\widehat{\lambda} - \lambda_a)_+$ . As a consequence, the restriction of  $(\widehat{W} - \widehat{W}_1)$  to  $\mathcal{R} \times \mathcal{R}^c$  is, up to a measure preserving bijection of its rows and of its columns, equal to the restriction of  $U$  to the set  $(\cup_{a: \lambda_a < \widehat{\lambda}_a} [F_{\widehat{\lambda}_\delta}(a - 1), F_{\widehat{\lambda}_\delta}(a)]) \times (\cup_a [F_\phi(a - 1), F_\phi(a - 1) + \widehat{\lambda}_a \wedge \lambda_a])$ . This implies that  $\|(\widehat{W} - \widehat{W}_1)|_{\mathcal{R} \times \mathcal{R}^c}\|_\square \leq \|U\|_\square$ . Thus, we only have to control  $\mathbb{E}[\|U\|_\square]$ .

**Step 4: Control of  $\mathbb{E}[\|U\|_\square]$ .** Define the sets  $\mathcal{B}_1 := \prod_{a=1}^k [0, |\widehat{\lambda}_a - \lambda_a|]$  and  $\mathcal{B}_2 := \prod_{a=1}^k [0, |\lambda_a|]$ . Then, the cut norm of  $U$  writes as

$$\begin{aligned} \|U\|_\square &\leq \sup_{\gamma \in \mathcal{B}_1, \gamma' \in \mathcal{B}_2} \left| \sum_{a,b=1}^k \gamma_a \gamma'_b (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)}) \right| \\ &\leq \sup_{S, T \in [k]} \left| \sum_{a \in S, b \in T} \lambda_b |\widehat{\lambda}_a - \lambda_a| (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)}) \right|, \end{aligned} \quad (49)$$

since the supremum of a linear function on a convex set is achieved at an extremal point. The random variable  $|\widehat{\lambda}_a - \lambda_a|$  is in expectation of the order  $\sqrt{\lambda_a/n}$ . If we could replace each  $|\widehat{\lambda}_a - \lambda_a|$  by  $\sqrt{\lambda_a/n}$  in (49), then thanks to (40), we could prove that  $\|U\|_\square$  is (up to a multiplicative constant) less than  $\sqrt{k/(n \log(k))}$ . Unfortunately, if we directly applied Bernstein inequality or the bounded difference inequality to simultaneously control  $|\widehat{\lambda}_a - \lambda_a|$  over all  $a \in [k]$  or to simultaneously control  $\sum_{a \in S, b \in T} \lambda_b |\widehat{\lambda}_a - \lambda_a| (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)})$  over all  $S, T \subset [k]$ , we would lose at least a logarithmic factor.

To bypass this issue, we adapt Lemma 10.9 of [28], which is a key point in the proof of sampling Lemma for graphons (Lemma 10.5 in [28]). Given a bounded non-symmetric kernel  $W \in \mathcal{W}_{\mathcal{X}, \mathcal{Y}}$ , let us define the following one-side version of the cut norm:

$$\|W\|_\square^+ = \sup_{X \subset \mathcal{X}, Y \subset \mathcal{Y}} \int_{X \times Y} W(x, y) dx dy,$$

where we take the supremum without any absolute value. As a consequence, the cut norm  $\|W\|_\square$  is the maximum  $\|W\|_\square^+$  and  $\| -W \|_\square^+$ .

**Lemma 6.** *Let  $W \in \mathcal{W}_{[0,u],[0,v]}[k]$  and let  $\mathbf{Q} \in \mathbb{R}^{k \times k}$ ,  $\phi_1 : [0, u] \rightarrow [k]$  and  $\phi_2 : [0, v] \rightarrow [k]$  be associated to  $W$  as in (31). For  $a = 1, \dots, k$ , define  $\alpha_a := \lambda(\phi_1^{-1}(\{a\}))$  and  $\beta_a := \lambda(\phi_2^{-1}(\{a\}))$ . Given any subset  $R \subset [k]$ , let*

$$R^{l,W} := \{b, \sum_{a \in R} \alpha_a \mathbf{Q}_{ab} > 0\}, \quad R^{r,W} := \{a, \sum_{b \in R} \beta_b \mathbf{Q}_{ab} > 0\}. \quad (50)$$

Finally, we define for any  $S, T \subset [k]$ ,  $W[S, T] := \sum_{a \in S, b \in T} \alpha_a \beta_b \mathbf{Q}_{ab}$ . Then, for any integer  $q$  with  $1 \leq q \leq k$ , we have

$$\|W\|_\square^+ \leq \max_{R_i \subset [k], |R_i| \leq q} W[R_2^{r,W}, R_1^{l,W}] + \frac{u \sqrt{k \sum_{a=1}^k \beta_a^2} + v \sqrt{k \sum_{a=1}^k \alpha_a^2}}{\sqrt{q}}. \quad (51)$$

Note that in contrast to Equation (49) where one considers a supremum of  $2^{2k}$  sums, only  $k^{2q}$  terms are involved in (51) up to the price of an additive term of order  $q^{-1/2}$ . The difficulty is that we will apply this lemma to  $U$  for which these  $k^{2q}$  will turn out to be random.

In the sequel, we fix  $q = \lfloor \sqrt{k} \rfloor$  and apply Lemma 6 to  $U$ . Then, we can take  $u = v = 1$ . Since  $\sum_{a=1}^k \lambda_a = 1$  and since we assumed at the beginning of the proof that the weights  $\lambda_a$  are all smaller than  $2/k$ , it follows that  $(k \sum_{a=1}^k \lambda_a^2)^{1/2} \leq \sqrt{2}$ . Let  $M$  and  $N$  denote the random variables  $M := \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a|$  and  $N := \left( \sum_{a=1}^k k |\hat{\lambda}_a - \lambda_a|^2 \right)^{1/2}$ . Both  $M$  and  $N$  are functions of the independent random variables  $(\xi_1, \dots, \xi_n)$ . Besides, if we change the values of one of these  $\xi_i$  the value of  $M$  changes by at most  $2/n$  and the value of  $N$  changes by at most  $\sqrt{2k}/n$ . As a consequence, we may apply the bounded difference inequality (Lemma (1)) to these two random variables. Then, with probability larger than  $1 - 2\exp(-\sqrt{k}/\log(k))$ , one has

$$\sum_{a=1}^k |\hat{\lambda}_a - \lambda_a| \leq \mathbb{E} \left[ \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a| \right] + \sqrt{\frac{2k^{1/2}}{n \log(k)}} \leq C \sqrt{\frac{k}{n}}, \quad (52)$$

$$\left( k \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a|^2 \right)^{1/2} \leq \mathbb{E} \left[ \left( k \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a|^2 \right)^{1/2} \right] + \sqrt{\frac{2k^{3/2}}{n \log(k)}} \leq C k^{1/4} \sqrt{\frac{k}{n \log(k)}}. \quad (53)$$

In (52) - (53) we bound the expectation using that, since  $\xi_1, \dots, \xi_n$  are i.i.d. uniformly distributed random variables,  $n\hat{\lambda}_a$  has a binomial distribution with parameters  $(n, \lambda_a)$  and the Cauchy-Schwarz inequality:

$$\begin{aligned} \mathbb{E} \left[ \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a| \right] &\leq \sum_{a=1}^k \sqrt{\frac{\lambda_a(1-\lambda_a)}{n}} \leq \sqrt{\frac{k}{n}} \quad \text{and} \\ \mathbb{E} \left[ \left( k \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a|^2 \right)^{1/2} \right] &\leq \sqrt{k \sum_{a=1}^k \frac{\lambda_a(1-\lambda_a)}{n}} \leq \sqrt{\frac{k}{n}}. \end{aligned}$$

Bound (53) and  $(k \sum_{a=1}^k \lambda_a^2)^{1/2} \leq \sqrt{2}$ , implies that for  $U$ , with probability larger than  $1 - 2\exp(-\sqrt{k}/\log(k))$ ,

$$\frac{\sqrt{k \sum_{a=1}^k \beta_a^2} + \sqrt{k \sum_{a=1}^k \alpha_a^2}}{k^{1/4}} \leq C \sqrt{\frac{k}{n \log(k)}}. \quad (54)$$

Fix any two subsets  $R_1, R_2 \subset [k]$  of size less than or equal to  $q$ . In view of (51), one needs to control the following random variable

$$Z_{R_1, R_2} := U \left[ R_2^{r, U}, R_1^{l, U} \right] = \sum_{a \in R_2^{r, U}} |\hat{\lambda}_a - \lambda_a| \sum_{b \in R_1^{l, U}} \lambda_b (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad}). \quad (55)$$

It is done in the following Lemma:

**Lemma 7.** *Let  $R_1, R_2$  be two subsets of  $[k]$  of size less than or equal to  $q$  and  $Z_{R_1, R_2}$  given by (55). Then, we have that with probability larger than  $1 - (1 + 2k)\exp(-\sqrt{k}/\log(k))$ ,*

$$\max_{R_1, R_2 : |R_1| \leq q, |R_2| \leq q} Z_{R_1, R_2} \leq C \sqrt{\frac{k}{n \log(k)}}.$$

Now, it follows from Lemma 6 together with (54) and Lemma 7 that, with probability larger than  $1 - (3 + 2k) \exp(-\sqrt{k}/\log(k))$ ,

$$\|U\|_{\square}^+ \leq C \sqrt{\frac{k}{n \log(k)}}.$$

Controlling analogously  $\| -U \|_{\square}^+$ , we conclude that there exists an event  $\mathcal{A}$  of probability larger than  $1 - 10 \exp(-\sqrt{k}/\log(k))$  such that, on  $\mathcal{A}$ ,

$$\|U\|_{\square} \leq C \sqrt{\frac{k}{n \log(k)}}.$$

To finish the control of  $\mathbb{E}[\|U\|_{\square}]$ , we use the rough bound  $\|U\|_{\square} \leq \|U\|_1 \leq \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a|$  on the complementary event  $\bar{\mathcal{A}}$ .

$$\begin{aligned} \mathbb{E}[\|U\|_{\square}] &\leq \mathbb{E}[\|U\|_{\square} \mathbf{1}_{\mathcal{A}}] + \mathbb{E}[\|U\|_{\square} \mathbf{1}_{\bar{\mathcal{A}}}] \\ &\leq C \sqrt{\frac{k}{n \log(k)}} + \sqrt{\mathbb{P}(\bar{\mathcal{A}})} \left[ \mathbb{E} \left( \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a| \right)^2 \right]^{1/2} \\ &\leq C \sqrt{\frac{k}{n \log(k)}} + C' e^{-\sqrt{k}/(2 \log(k))} \frac{k}{\sqrt{n}} \leq C'' \sqrt{\frac{k}{n \log(k)}} \end{aligned} \quad (56)$$

where we use (52). Now, using the decomposition (45), (46) and (48), we can conclude that

$$\mathbb{E} \left[ \left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}^c} \right\|_{\square} \right] \leq C \sqrt{\frac{k}{n \log(k)}}.$$

The following lemma gives a corresponding bound on the second term  $\left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}} \right\|_{\square}$  in (44). The proof is somewhat analogous to that of the control of  $\left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}^c} \right\|_{\square}$  and is deferred to the end of the section.

**Lemma 8.** *We have*

$$\mathbb{E} \left[ \left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}} \right\|_{\square} \right] \leq C \sqrt{\frac{k}{n \log(k)}}.$$

In view of (44), we have proved Theorem 1.  $\square$

*Proof of Lemma 5.* For  $a \in [k]$ , we denote  $(\lambda_0)_a = \lambda(\phi_0^{-1}(a))$  and  $u_a = \frac{\sqrt{(\lambda_0)_a}}{\sum_b \sqrt{(\lambda_0)_b}}$ . For any  $b \in [k]$ , define the cumulative distribution functions  $F_0(b) = \sum_{a=1}^b (\lambda_0)_a$  and  $F_1(b) = \sum_{a=1}^b u_a$ . For  $a, b \in [k]$ , let  $(\Pi_d)_{ab} = [F_0(a-1), F_0(a)) \times [F_1(b-1), F_1(b))$  and  $(\Pi_d^+)_{ab} = [F_1(a-1), F_1(a)) \times [F_1(b-1), F_1(b))$ . Finally we consider the (non necessarily symmetric) kernels  $W_d$  and  $W_d^+$  defined by

$$W_d(x, y) = \sum_{a=1}^k \sum_{b=1}^k (\mathbf{Q}_0)_{ab} \mathbf{1}_{(\Pi_d)_{ab}}(x, y), \quad W_d^+(x, y) = \sum_{a=1}^k \sum_{b=1}^k (\mathbf{Q}_0)_{ab} \mathbf{1}_{(\Pi_d^+)_{ab}}(x, y).$$

In comparison to  $W_0$ , the length of the steps in  $W_d$  and  $W_d^+$  has been modified.

**Lemma 9.** Let  $W \in \mathcal{W}_{[0,1],[0,1]}$  be a  $k$ -step kernel defined by

$$W(x, y) = \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \mathbb{1}_{S_a \times T_b}(x, y)$$

where  $\mathbf{Q} \in [0, 1]^{k \times k}$  and  $(S_1, \dots, S_k)$  and  $(T_1, \dots, T_k)$  are two partitions of  $[0, 1]$  into a finite number of measurable sets. For any integer  $q_0 \geq 2$ , there exist a  $q_0$ -step kernel  $W^{(ap)} \in \mathcal{W}_{[0,1],[0,1]}^+$  satisfying

(i) for any  $(a, b) \in [k]$ ,  $W^{(ap)}$  is constant on  $S_a \times T_b$  and

(ii)  $\|W - W^{(ap)}\|_{\square} \leq \frac{C}{\sqrt{\log(q_0)}}.$

The second property (ii) is just the consequence of the weak Regularity Lemma for kernels [19] (see also Corollary 9.13 in [28]). The first property, (i), follows from the explicit construction of the approximate kernel by Kannan and Frieze (see the proof of Lemma 9.10 in [28]). For the sake of completeness, we give the details in the end of this section.

Fix  $q_0 = \lfloor k^{1/4} \rfloor$ . Note that  $q_0 \geq 2$  since we assume that  $k \geq 16$ . We denote by  $W_d^{(ap)}$  and  $W_d^{(ap,+)}$  the  $q_0$ -step kernels given by Lemma 9 to respectively approximate  $W_d$  and  $W_d^{(+)}$ . In virtue of Property (i), there exist two matrices  $\mathbf{Q}_0^{(ap)}$  and  $\mathbf{Q}_0^{(ap,+)}$  in  $[0, 1]^{k \times k}$  such that

$$W_d^{(ap)}(x, y) = \sum_{a=1}^k \sum_{b=1}^k (\mathbf{Q}_0^{(ap)})_{ab} \mathbb{1}_{(\Pi_d)_{ab}}(x, y) \quad \text{and} \quad W_d^{(ap,+)}(x, y) = \sum_{a=1}^k \sum_{b=1}^k (\mathbf{Q}_0^{(ap,+)})_{ab} \mathbb{1}_{(\Pi_d^+)_{ab}}(x, y).$$

There exist two partitions  $\mathcal{P}_d$  and  $\mathcal{P}_d^+$  of  $[k]$  such that  $\mathbf{Q}_0^{(ap)}$  is block constant according to  $\mathcal{P}_d$  and  $\mathbf{Q}_0^{(ap,+)}$  is block constant according to  $\mathcal{P}_d^+$ . Let  $\mathcal{P}^*$  be the coarsest partition that refines both  $\mathcal{P}$  and  $\mathcal{P}_d^+$ . As a consequence,  $\mathcal{P}^*$  is made of less than  $q_0^2 \leq q_1$  subsets. By possibly refining  $\mathcal{P}^*$ , we may assume without loss of generality that  $\mathcal{P}^* = (P_1^*, \dots, P_{q_1}^*)$  is made of exactly  $q_1$  elements. Let  $\pi$  be a permutation of  $[k]$  transforming  $\mathcal{P}^*$  in a partition  $\mathcal{P} = (P_1, \dots, P_{q_1})$  with  $P_a = \{\pi(b), b \in P_a^*\}$  made of consecutive intervals. Denoting  $\Pi$  the corresponding permutation matrix, we finally take

$$\mathbf{Q} = \Pi^T \mathbf{Q}_0 \Pi, \quad \mathbf{Q}^{(ap)} = \Pi^T \mathbf{Q}_0^{(ap)} \Pi, \quad \text{and} \quad \mathbf{Q}^{(ap,+)} = \Pi^T \mathbf{Q}_0^{(ap,+)} \Pi.$$

Now we are ready to prove (40) and (41). Recall that we denote  $\phi = \pi \circ \phi_0$  and  $\lambda_a := \lambda(\phi^{-1}(a))$  for  $a \in [k]$ . Define the sets  $\mathcal{B}_1 := \prod_{a=1}^k [0, u_{\pi(a)}]$  and  $\mathcal{B}_2 := \prod_{a=1}^k [0, \lambda_a]$ . Since  $W_d - W_d^{(ap)}$  is a  $k$ -step function, its cut norm writes as

$$\|W_d - W_d^{(ap)}\|_{\square} = \sup_{\gamma \in \mathcal{B}_1, \gamma \in \mathcal{B}_2} \left| \sum_{a,b} \gamma_a \gamma_b (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)}) \right| \quad (57)$$

$$= \sup_{\epsilon \in \{0,1\}^k, \epsilon' \in \{0,1\}^k} \left| \sum_{a,b} \epsilon_a \epsilon'_b \lambda_b u_{\pi(a)} (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)}) \right| \leq \frac{C}{\sqrt{\log(q_0)}} \quad (58)$$

since the supremum is achieved at an extremal point of the convex and in the last inequality we use property (ii) of Lemma 9. Now (57) and the definition of  $u_{\pi(a)}$  imply

$$\sup_{\epsilon \in \{0,1\}^k, \epsilon' \in \{0,1\}^k} \left| \sum_{a,b} \epsilon_a \epsilon'_b \lambda_b \sqrt{\lambda_a} (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap)}) \right| \leq C \frac{\sum_{b \in [k]} \sqrt{\lambda_b}}{\sqrt{\log(q_0)}} \leq C' \sqrt{\frac{k}{\log(k)}},$$

by Cauchy-Schwarz inequality. We have proved (40). The second inequality (41) is derived similarly.  $\square$

**Proof of Lemma 9.** We adapt the proof of the weak Regularity Lemma for symmetric kernels [28, Lemma 9.9] to non symmetric ones. We use the following extension of Lemma 9.11(a) in [28].

**Lemma 10.** *For every  $W \in \mathcal{W}_{[0,1],[0,1]}[k]$  such that*

$$W(x, y) = \sum_{a=1}^k \sum_{b=1}^k Q_{ab} \mathbf{1}_{S_a \times T_b}(x, y)$$

where  $Q \in \mathbb{R}^{k \times k}$  and  $\mathcal{P} = \{(S_1, \dots, S_k), (T_1, \dots, T_k)\}$  are two partitions of  $[0, 1]$  into a finite number of measurable sets, there are two sets  $\mathcal{A}, \mathcal{B} \subset [k]$  and a real number  $0 \leq a \leq \max_{a,b} |Q_{ab}|$  such that, for  $S' = \cup_{a \in \mathcal{A}} S_a$  and  $T' = \cup_{b \in \mathcal{B}} T_b$ ,

$$\|W - a \mathbf{1}_{S' \times T'}\|_2^2 \leq \|W\|_2^2 - \|W\|_{\square}^2.$$

Now we apply Lemma 10 repeatedly, to get pairs of sets  $S'_i, T'_i$  and real numbers  $a_i$  such that for any positive integer  $j$ ,  $W_j = W - \sum_{i=1}^j a_i \mathbf{1}_{S'_i \times T'_i}$  we have

$$\|W_j\|_2^2 \leq \|W\|_2^2 - \sum_{i=1}^{j-1} \|W_i\|_{\square}^2.$$

Fix some integer  $k_0 > 0$ . Since the right-hand side of the above equation remains non-negative, there exists  $0 \leq i < k_0$  with  $\|W_i\|_{\square}^2 \leq 1/k_0$ . Now putting  $a_l = 0$  for  $l > i$  we get that for any  $W \in \mathcal{W}_{[0,1],[0,1]}[k]$  and any  $k_0 \geq 1$  there are  $k_0$  pairs of subsets  $S'_i, T'_i \subset [0, 1]$  and  $k_0$  real numbers  $a_i$  such that

$$\left\| W - \sum_{i=1}^{k_0} a_i \mathbf{1}_{S'_i \times T'_i} \right\|_{\square} < \frac{1}{\sqrt{k_0}}. \quad (59)$$

Note that the approximation  $W^{ap} = \sum_{i=1}^{k_0} a_i \mathbf{1}_{S'_i \times T'_i}$  is a step function with at most  $2^{k_0}$  steps and  $a_i \geq 0$ , for all  $i$ . On the other hand, by construction we have that for any  $(a, b) \in [k]$ ,  $W^{(ap)}$  is constant on all sets of the form  $S_a \times T_b$ . We conclude by taking  $k_0 = \lfloor \log(q_0)/\log(2) \rfloor$ .  $\square$

**Proof of Lemma 10.** This lemma is proved in [28, Lemma 9.11] for symmetric kernels. For readers convenience we get the details here. Let  $W$  be a  $k$ -step kernel and let  $(S_1, \dots, S_k), (T_1, \dots, T_k)$  be two measurable partitions of  $[0, 1]$  such that  $W$  is constant on each set  $S_i \times T_j$ . Relying on a convexity argument as in the proof of Lemma 5, the cut norm is achieved for measurable sets  $S$  and  $T$  that are unions of  $S_i$  and  $T_j$  respectively, that is

$$\|W\|_{\square} = \left| \int_{S \times T} W(x, y) dx dy \right|,$$

where  $S = \cup_{a \in \mathcal{A}} S_a$  and  $T = \cup_{b \in \mathcal{B}} T_b$  with  $\mathcal{A}, \mathcal{B} \subset [k]$ . Let  $\mathbf{a} = \frac{1}{\lambda(S)\lambda(T)} \|W\|_{\square}$ . Then, we have

$$\|W - \mathbf{a} \mathbf{1}_{S \times T}\|_2^2 = \|W\|_2^2 - \frac{1}{\lambda(S)\lambda(T)} \|W\|_{\square}^2 \leq \|W\|_2^2 - \|W\|_{\square}^2$$

which completes the proof.  $\square$

*Proof of Lemma 6.* This proof follows closely that of Lemma 10.9 in [28]. It is easy to see that

$$\|W\|_{\square}^+ = \max_{S, T \subset [k]} W[S, T]$$



so we only need to bound these expressions. Let  $Q$  and  $Q'$  be independent uniformly chosen  $q$ -subset of  $[k]$  and let  $\mathbb{E}_Q$  (resp.  $\mathbb{E}_{Q'}$ ) denote the expectation with respect to  $Q$  (resp.  $Q'$ ). We shall prove that, for any  $S, T \subset [k]$

$$W[S, T] \leq \mathbb{E}_Q [W[(Q \cap T)^{r,W}, T]] + \frac{u\sqrt{k \sum_{a=1}^k \beta_a^2}}{\sqrt{q}}. \quad (60)$$

By symmetry, this will imply

$$W[S, T] \leq \mathbb{E}_{Q'} [W[S, (Q' \cap S)^{l,W}]] + \frac{v\sqrt{k \sum_{a=1}^k \alpha_a^2}}{\sqrt{q}},$$

so that gathering both inequalities yields to

$$W[S, T] \leq \mathbb{E}_{Q, Q'} [W[(Q \cap T)^{r,W}, (Q' \cap (Q \cap T)^{r,W})^{l,W}]] + \frac{u\sqrt{k \sum_{a=1}^k \beta_a^2} + v\sqrt{k \sum_{a=1}^k \alpha_a^2}}{\sqrt{q}}.$$

Since the above expectation is less than or equal to  $\sup_{R_i, |R_i| \leq q} W[R_2^{r,W}, R_1^{l,W}]$ , this will conclude the proof. Thus, we only have to show (60). Note that  $W[S, T] \leq W[T^{r,W}, T]$  implies that it suffices to prove

$$\mathbb{E}_Q [W[T^{r,W} \setminus (Q \cap T)^{r,W}, T]] - \mathbb{E}_Q [W[(Q \cap T)^{r,W} \setminus T^{r,W}, T]] \leq \frac{u\sqrt{k \sum_{a=1}^k \beta_a^2}}{\sqrt{q}}. \quad (61)$$

Let us denote  $Z$  the above difference of expectations. For any  $a \in [k]$ , write  $B_a = \sum_{b \in T} \beta_b \mathbf{Q}_{ab}$  and  $A_a = \sum_{b \in T \cap Q} \beta_b \mathbf{Q}_{ab}$ . By the definition (50), we have that  $B_a$  is non-negative for  $a \in T^{r,W}$  and  $B_a \leq 0$  if  $a \notin T^{r,W}$ . In the same way,  $A_a > 0$  for  $a \in (Q \cap T)^{r,W}$  and  $A_a \leq 0$  for  $a \notin (Q \cap T)^{r,W}$ . Denoting  $\mathbb{P}_Q$  the probability with respect to  $Q$ , we obtain

$$\begin{aligned} Z &= \mathbb{E}_Q \left( \sum_{a \in T^{r,W}} \mathbb{1}_{\{a \notin (Q \cap T)^{r,W}\}} \alpha_a B_a + \sum_{a \notin T^{r,W}} \mathbb{1}_{\{a \in (Q \cap T)^{r,W}\}} \alpha_a |B_a| \right) \\ &= \sum_{a \in T^{r,W}} \mathbb{P}_Q[A_a \leq 0] \alpha_a B_a + \sum_{a \notin T^{r,W}} \mathbb{P}_Q[A_a > 0] \alpha_a |B_a|. \end{aligned} \quad (62)$$

Now, using  $\mathbb{E}_Q[A_a] = qB_a/k$ , it follows from the Chebyshev inequality that, for  $a \in T^{r,W}$ , we have  $\mathbb{P}_Q[A_a < 0] \leq \text{Var}_Q[A_a]/\mathbb{E}_Q^2[A_a]$ . Since a probability is smaller or equal to one, it follows that  $\mathbb{P}_Q[A_a < 0] \leq \sqrt{\text{Var}_Q[A_a]}/|\mathbb{E}_Q[A_a]|$ . Similarly, for  $a \notin T^{r,W}$  we also have that  $\mathbb{P}_Q[A_a > 0] \leq \sqrt{\text{Var}_Q[A_a]}/|\mathbb{E}_Q[A_a]|$ . Coming back to  $Z$ , this yields

$$Z \leq \sum_{a \in [k]} \alpha_a |B_a| \frac{\text{Var}_Q^{1/2}[A_a]}{|\mathbb{E}_Q[A_a]|} = \frac{k}{q} \sum_{a \in [k]} \alpha_a \text{Var}_Q^{1/2}[A_a] \leq \frac{ku}{q} \max_{a \in [k]} \text{Var}_Q^{1/2}[A_a].$$

Working out the variance, we get  $\text{Var}_Q[A_a] \leq \frac{q}{k} \sum_{b \in T} \beta_b^2 \mathbf{Q}_{ab}^2 \leq q(\sum_{b \in [k]} \beta_b^2)/k$ , which concludes the proof.  $\square$

*Proof of Lemma 7.* Note that in (55), the definition of  $Z_{R_1, R_2}$ , the set  $R_2^{r, U}$  is deterministic whereas the set  $R_1^{l, U}$  only depends on  $(\hat{\lambda}_a)_{a \in R_1}$ . We can upper bound  $Z_{R_1, R_2}$  in the following way:

$$Z_{R_1, R_2} \leq \sum_{a \in R_2^{r, U} \setminus R_1} \left| \hat{\lambda}_a - \lambda_a \right| \sum_{b \in R_1^{l, U}} \lambda_b \left( \mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad} \right) + \sum_{a \in R_1} \left| \hat{\lambda}_a - \lambda_a \right| \quad (63)$$

where we use  $\left| \sum_b \lambda_b (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad}) \right| \leq 1$ . We set

$$T_{R_1, R_2} = \sum_{a \in R_2^{r, U} \setminus R_1} \left| \hat{\lambda}_a - \lambda_a \right| \sum_{b \in R_1^{l, U}} \lambda_b \left( \mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad} \right).$$

Conditionally to  $(\hat{\lambda}_a)_{a \in R_1}$ ,  $T_{R_1, R_2}$  is distributed as a function of  $n - n \sum_{a \in R_1} \hat{\lambda}_a$  i.i.d. random variables  $\xi'_i$  such that  $\mathbb{P}[\xi' = a] = \lambda_a / (1 - \sum_{a \in R_1} \lambda_a)$  for any  $a \in [k] \setminus R_1$ . Besides, if we change the values of one of these  $\xi'_i$  the value of this expression changes by at most  $2/n$ . It then follows from the bounded difference inequality (Lemma (1)) that, for any  $t > 0$

$$\mathbb{P} \left\{ T_{R_1, R_2} \geq \mathbb{E} [T_{R_1, R_2} | (\hat{\lambda}_a)_{a \in R_1}] + \sqrt{\frac{2t}{n}} | (\hat{\lambda}_a)_{a \in R_1} \right\} \geq 1 - e^{-t}. \quad (64)$$

Let us bound this conditional expectation:

$$\begin{aligned} \mathbb{E} [T_{R_1, R_2} | (\hat{\lambda}_a)_{a \in R_1}] &= \sum_{a \in R_2^{r, U} \setminus R_1} \mathbb{E} \left[ \left| \hat{\lambda}_a - \lambda_a \right| | (\hat{\lambda}_c)_{c \in R_1} \right] \sum_{b \in R_1^{l, U}} \lambda_b \left( \mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad} \right) \\ &\leq \sup_{S \subset [k] \setminus R_1, T \subset [k]} \sum_{a \in S, b \in T} \mathbb{E} \left[ \left| \hat{\lambda}_a - \lambda_a \right| | (\hat{\lambda}_c)_{c \in R_1} \right] \lambda_b (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad}). \end{aligned} \quad (65)$$

Now, using Cauchy-Schwarz inequality, we have

$$\mathbb{E} \left[ \left| \hat{\lambda}_a - \lambda_a \right| | (\hat{\lambda}_c)_{c \in R_1} \right] \leq \sqrt{\frac{\lambda_a}{n(1 - \sum_{b \in R_1} \lambda_b)}} \leq \sqrt{\frac{\lambda_a}{n(1 - 2q/k)}} \leq 2\sqrt{\frac{\lambda_a}{n}},$$

where we used that  $\lambda_b \leq 2/k$ ,  $|R_1| \leq q \leq k^{1/2}$  and  $k \geq 8$ . The supremum in (65) is achieved for subsets  $(S^*, T^*)$  such that for all  $a \in S^*$ ,  $\sum_{b \in T^*} \lambda_b (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad})$  is non-negative (otherwise this contradicts the optimality of  $S^*, T^*$ ). As a consequence, we can plug the upper bounds on  $\mathbb{E} \left[ \left| \hat{\lambda}_a - \lambda_a \right| | (\hat{\lambda}_c)_{c \in R_1} \right]$  into (65):

$$\mathbb{E} [T_{R_1, R_2} | (\hat{\lambda}_a)_{a \in R_1}] \leq \frac{2}{\sqrt{n}} \sup_{S \subset [k] \setminus R_1, T \subset [k]} \sum_{a \in S, b \in T} \sqrt{\lambda_a} \lambda_b (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad}) \leq C \sqrt{\frac{k}{n \log(k)}},$$

where we used the property (40) of  $\mathbf{Q}^{ad}$ . Coming back to (64) and integrating the deviation inequality with respect to  $(\hat{\lambda}_a)_{a \in R_1}$ , we conclude that, for any  $t > 0$

$$\mathbb{P} \left[ T_{R_1, R_2} \geq C \sqrt{\frac{k}{n \log(k)}} + \sqrt{\frac{2t}{n}} \right] \geq 1 - e^{-t}.$$

Fixing  $t = 2 \log(k)q + \sqrt{k}/\log(k)$  and taking an union bound over all possible  $R_1, R_2$ , we derive that

$$\max_{R_1, R_2, |R_1| \leq q, |R_2| \leq q} Z_{R_1, R_2} \leq C \sqrt{\frac{k}{n \log(k)}} + q \max_{a=1, \dots, k} \left| \hat{\lambda}_a - \lambda_a \right| \quad (66)$$

on an event of probability higher than  $1 - \exp(-\sqrt{k}/\log(k))$ .

Next we bound  $\max_{a=1,\dots,k} |\hat{\lambda}_a - \lambda_a|$ . Recall that  $n\hat{\lambda}_a$  has a binomial distribution with parameters  $(n, \lambda_a)$  and  $\lambda_a \leq 2/k$ . For any  $a \in [k]$ , applying Bernstein inequality to  $|\hat{\lambda}_a - \lambda_a|$  we get

$$\mathbb{P} \left\{ n|\hat{\lambda}_a - \lambda_a| \geq t \right\} \leq 2 \exp \left( -\frac{t^2}{4n/k + 2t/3} \right).$$

Taking  $t = C\sqrt{n/\log(k)}$  (for a suitable constant  $C > 0$ ) and applying the union bound, we derive that with probability larger than  $1 - 2k \exp(-\sqrt{k}/\log(k))$

$$\sqrt{k} \max_{a=1,\dots,k} |\hat{\lambda}_a - \lambda_a| \leq C\sqrt{k/(n \log(k))}. \quad (67)$$

The bound (66) together with (67) imply the statement of Lemma 7.  $\square$

*Proof of Lemma 8.* As the control of  $(W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}}$  is quite similar to that of  $(W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}^c}$ , we only sketch the main steps. Relying on the graphon  $W_1^+$  (defined in (42)), we have the following decomposition:

$$\|(W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}}\|_{\square} \leq \|(W_1^+ - \widehat{W}_1^+)|_{\mathcal{R} \times \mathcal{R}}\|_{\square} + \|(W - W_1^+)|_{\mathcal{R} \times \mathcal{R}}\|_{\square} + \|(\widehat{W} - \widehat{W}_1^+)|_{\mathcal{R} \times \mathcal{R}}\|_{\square}. \quad (68)$$

Since  $(W_1^+ - \widehat{W}_1^+)(x, y)$  is zero except if  $x \in \mathcal{R}_2$  or  $y \in \mathcal{R}_2$ , we bound the first expression by its  $l_1$  norm as for  $W_1 - \widehat{W}_1$ :

$$\mathbb{E} [\|(W_1^+ - \widehat{W}_1^+)|_{\mathcal{R} \times \mathcal{R}}\|_{\square}] \leq 2\mathbb{E}[\lambda(\mathcal{R}_2)] \leq 2\frac{k^{1/4}}{\sqrt{n}}. \quad (69)$$

The two last expressions in (68) are bounded by the cut norm of a kernel  $V$  defined as follows. For any  $a, b \in [k]$ , define  $\tilde{\Pi}_{a,b} = [F_{\hat{\lambda}^\delta}(a-1), F_{\hat{\lambda}^\delta}(a)] \times [F_{\hat{\lambda}^\delta}(b-1), F_{\hat{\lambda}^\delta}(b)]$  where  $F_{\hat{\lambda}^\delta}(\cdot)$  has been defined in (47). Let  $V$  be the  $k \times k$  step kernel on  $[0, \sum_a |\hat{\lambda}_a - \lambda_a|]^2$  given by

$$V(x, y) := \sum_{a,b=1}^k [\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap,+)}] \mathbf{1}_{\tilde{\Pi}_{a,b}}(x, y).$$

Now, as for the restrictions of  $W - W_1$  and  $\widehat{W} - \widehat{W}_1$  to  $\mathcal{R} \times \mathcal{R}^c$ , we have

$$\|(W - W_1^+)|_{\mathcal{R} \times \mathcal{R}}\|_{\square} \vee \|(\widehat{W} - \widehat{W}_1^+)|_{\mathcal{R} \times \mathcal{R}}\|_{\square} \leq \|V\|_{\square}. \quad (70)$$

Thus, it boils down to controlling  $\mathbb{E}[\|V\|_{\square}]$ . Since  $V$  is a  $k$ -step kernel, its cut norm writes as

$$\|V\|_{\square} = \sup_{S, T \subset [k]} \left| \sum_{a \in S, b \in T} |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{(ap,+)}) \right|.$$

As for the kernel  $U$  in the main proof, we rely on the Lemma 6. The random variables  $\sum_a |\hat{\lambda}_a - \lambda_a|$  and  $(\sum_a |\hat{\lambda}_a - \lambda_a|^2)^{1/2}$  are controlled as in (52) and (53).

Fix any two subsets  $R_1, R_2 \subset [k]$  of size less than or equal to  $q$  and define

$$Z_{R_1, R_2} := V[R_2^{r,V}, R_1^{l,V}] = \sum_{a \in R_2^{r,V}} \sum_{b \in R_1^{l,V}} |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}).$$

The set  $R_1^{l,V}$  only depends on  $(\hat{\lambda}_a)_{a \in R_1}$  and  $R_2^{r,V}$  only depends on  $(\hat{\lambda}_a)_{a \in R_2}$ . We have

$$Z_{R_1, R_2} \leq \sum_{a \in R_2^{r,V} \setminus (R_1 \cup R_2)} \sum_{b \in R_1^{l,V} \setminus (R_1 \cup R_2)} |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}) + 4 \sum_{a \in R_1 \cup R_2} |\hat{\lambda}_a - \lambda_a| ,$$

since  $\sum_{a \in [k]} |\hat{\lambda}_a - \lambda_a| \leq 2$ . We set

$$T_{R_1, R_2} = \sum_{a \in R_2^{r,V} \setminus (R_1 \cup R_2)} \sum_{b \in R_1^{l,V} \setminus (R_1 \cup R_2)} |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}).$$

Write  $R := R_1 \cup R_2$  and  $\hat{\lambda}_{\{R\}} := (\hat{\lambda}_a)_{a \in R}$ . Conditionally to  $\hat{\lambda}_{\{R\}}$ ,  $T_{R_1, R_2}$  is a function of  $n - n \sum_{a \in R} \hat{\lambda}_a$  independent random variables. Besides, if we change the values of one of these independent random variables the value of  $T_{R_1, R_2}$  changes by at most  $4/n$ . Hence, the bounded difference inequality enforces that, for any  $t > 0$ ,

$$\mathbb{P} \left[ T_{R_1, R_2} \geq \mathbb{E}[T_{R_1, R_2} | \hat{\lambda}_{\{R\}}] + 8 \sqrt{\frac{2t}{n}} |\hat{\lambda}_{\{R\}}| \right] \geq 1 - e^{-t}. \quad (71)$$

The conditional expectation is upper bounded by

$$\mathbb{E}[T_{R_1, R_2} | \hat{\lambda}_{\{R\}}] \leq \sup_{S \subset [k] \setminus R, T \subset [k] \setminus R} \sum_{a \in S, b \in T} \mathbb{E} [ |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| | \hat{\lambda}_{\{R\}} ] (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}). \quad (72)$$

Here, unfortunately, we cannot directly replace  $\mathbb{E} [ |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| | \hat{\lambda}_{\{R\}} ]$  by an upper bound of it because this expression does not factorize. We shall prove that  $\mathbb{E} [ |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| | \hat{\lambda}_{\{R\}} ]$  is, up to a small loss, close to a product of expectations.

Write  $N := n - n \sum_{c \in R} \hat{\lambda}_c$ ,  $\lambda_R := \sum_{c \in R} \lambda_c$  and  $\hat{\lambda}_R = \sum_{c \in R} \hat{\lambda}_c$ . Note that  $n \hat{\lambda}_R$  has a binomial distribution with parameters  $(n, \lambda_R)$ . Applying Bernstein inequality to  $|\hat{\lambda}_R - \lambda_R|$  we get

$$\mathbb{P} \left\{ n |\hat{\lambda}_R - \lambda_R| \geq t \right\} \leq 2 \exp \left( - \frac{t^2}{4n/\sqrt{k} + 2t/3} \right). \quad (73)$$

Let  $\mathcal{R} = \left\{ |\hat{\lambda}_R - \lambda_R| \leq \frac{1}{\sqrt{n \log(k)}} \right\}$ . Taking  $t = \sqrt{n/\log(k)}$  in (73) we have that

$$\mathbb{P}(\mathcal{R}) \geq 1 - 2e^{-\sqrt{k}/\log(k)}.$$

In what follows we assume that the event  $\mathcal{R}$  is true. Take any two distinct elements  $a$  and  $b$  of  $[k] \setminus R$ . We shall prove that the conditional expectations  $\mathbb{E} [ |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| | \hat{\lambda}_{\{R\}} ]$  are close to the products  $\mathbb{E} [ |\hat{\lambda}_a - \lambda_a| | \hat{\lambda}_{\{R\}} ] \mathbb{E} [ |\hat{\lambda}_b - \lambda_b| | \hat{\lambda}_{\{R\}} ]$ . It is easy to see that conditionally on  $(\hat{\lambda}_{\{R\}}, \hat{\lambda}_a)$ ,  $n \hat{\lambda}_b$  follows the Binomial distribution with parameters  $((N - n \hat{\lambda}_a), \lambda_b/(1 - \lambda_R - \lambda_a))$ . On the other hand, conditionally on  $\hat{\lambda}_{\{R\}}$ ,  $n \hat{\lambda}_b$  follows the Binomial distribution with parameters  $(N, \lambda_b/(1 - \lambda_R))$ . Let  $z_1, z_2, \dots$ , be a sequence of independent Bernoulli random variables with parameters  $\lambda_b/(1 - \lambda_a - \lambda_R)$ ,  $w_1, w_2, \dots$ , be an independent sequence of Bernoulli random variables with parameters  $(1 - \lambda_a - \lambda_R)/(1 - \lambda_R)$  and  $v_1, v_2, \dots$ , be an independent sequence of Bernoulli random variables with parameters  $\lambda_b/(1 - \lambda_R)$ . We define the following random variables:

$$X := \sum_{i=1}^{N-n\hat{\lambda}_a} z_i, \quad Y := \sum_{i=1}^{N-n\hat{\lambda}_a} z_i w_i + \sum_{i=1}^{n\hat{\lambda}_a} v_i$$

where we use  $\lambda_c \leq 2/k$  and  $|R| \leq 2\sqrt{k}$ . It is easy to see that  $X$  follows the Binomial distribution with parameters  $(N - n\hat{\lambda}_a)$  and  $\lambda_b/(1 - \lambda_R - \lambda_a)$  and  $Y$  follows the Binomial distribution with parameters  $N$  and  $\lambda_b/(1 - \lambda_R)$ . Hence, we have that

$$\left| \mathbb{E} [|\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| |\hat{\lambda}_{\{R\}}] - \mathbb{E} [|\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}}] \mathbb{E} [|\hat{\lambda}_b - \lambda_b| |\hat{\lambda}_{\{R\}}] \right| \leq \frac{1}{n} \mathbb{E} [|X - Y| |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}}]. \quad (74)$$

Relying our coupling between  $X$  and  $Y$ , we obtain

$$\begin{aligned} \frac{1}{n} \mathbb{E} [|X - Y| |\hat{\lambda}_{\{R\}}, \hat{\lambda}_a] &\leq \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^{N-n\hat{\lambda}_a} z_i (1 - \omega_i) \middle| \hat{\lambda}_{\{R\}}, \hat{\lambda}_a \right] + \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^{n\hat{\lambda}_a} v_i \middle| \hat{\lambda}_{\{R\}}, \hat{\lambda}_a \right] \\ &= \frac{N - n\hat{\lambda}_a}{n} \frac{\lambda_b \lambda_a}{(1 - \lambda_R)(1 - \lambda_a - \lambda_R)} + \hat{\lambda}_a \frac{\lambda_b}{1 - \lambda_R} \\ &\leq \frac{\lambda_b \lambda_a}{(1 - \lambda_R)(1 - \lambda_a - \lambda_R)} + \frac{\hat{\lambda}_a \lambda_b}{1 - \lambda_R}. \end{aligned} \quad (75)$$

On the other hand, conditionally on  $\hat{\lambda}_{\{R\}}$ ,  $n\hat{\lambda}_a$  follows the Binomial distribution with parameters  $(N, \lambda_a/(1 - \lambda_R))$  so that Cauchy-Schwarz inequality implies

$$\begin{aligned} \mathbb{E} [|\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}}] &= \mathbb{E} \left[ \left| \hat{\lambda}_a - \frac{N\lambda_a}{(1 - \lambda_R)n} + \frac{N\lambda_a}{(1 - \lambda_R)n} - \lambda_a \right| |\hat{\lambda}_{\{R\}} \right] \\ &\leq \mathbb{E} \left[ \left| \hat{\lambda}_a - \frac{N\lambda_a}{(1 - \lambda_R)n} \right| |\hat{\lambda}_{\{R\}} \right] + \left| \frac{N\lambda_a}{(1 - \lambda_R)n} - \lambda_a \right| \\ &\leq C\sqrt{\lambda_a/n} + \lambda_a |\lambda_R - \hat{\lambda}_R| \\ &\leq C\sqrt{\lambda_a/n} + \frac{4}{\sqrt{kn \log(k)}} \end{aligned} \quad (76)$$

where we use that  $\lambda_a \leq 2/k$  and the definition of the event  $\mathcal{R}$ . Similarly we compute

$$\mathbb{E} [\hat{\lambda}_a |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}}] \leq C \left( \frac{1}{kn} + \frac{1}{k\sqrt{kn}} \right) \quad (77)$$

Plugging (75 – 77) into (74) we get

$$\begin{aligned} &\left| \mathbb{E} [|\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| |\hat{\lambda}_{\{R\}}] - \mathbb{E} [|\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}}] \mathbb{E} [|\hat{\lambda}_b - \lambda_b| |\hat{\lambda}_{\{R\}}] \right| \\ &\leq \mathbb{E} \left[ \frac{\lambda_b \lambda_a}{(1 - \lambda_R)(1 - \lambda_a - \lambda_R)} |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}} \right] + \mathbb{E} \left[ \hat{\lambda}_a \frac{\lambda_b}{1 - \lambda_R} |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}} \right] \\ &\leq C \left[ \frac{1}{k^{5/2} n^{1/2}} + \frac{1}{nk^2} \right], \end{aligned}$$

where we use  $\lambda_b, \lambda_a \leq 2/k$ . For  $a = b$ , (76) implies that the above difference is of order  $(kn)^{-1}$ . Going back to (72), we obtain that

$$\mathbb{E} [T_{R_1, R_2} |\hat{\lambda}_{\{R\}}] \leq \sup_{S \subset [k] \setminus R, T \subset [k] \setminus R} \sum_{a \in S, b \in T} \mathbb{E} [|\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_{\{R\}}] \mathbb{E} [|\hat{\lambda}_b - \lambda_b| |\hat{\lambda}_{\{R\}}] (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}) + \frac{C}{\sqrt{n}}.$$

Take  $S^*$  and  $T^*$  being two sets maximizing the above expression. Then, for all  $a \in S^*$  we have that  $\sum_{b \in T^*} \mathbb{E} [|\hat{\lambda}_b - \lambda_b| |\hat{\lambda}_R|] (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+})$  is non-negative. As a consequence, using (76), we have that

$$\mathbb{E} [T_{R_1, R_2} | \hat{\lambda}_{\{R\}}] \leq C \sup_{S \subset [k] \setminus R, T \subset [k] \setminus R} \sum_{a \in S, b \in T} \left( \sqrt{\frac{\lambda_a}{n}} + \frac{4}{\sqrt{kn \log(k)}} \right) \mathbb{E} [|\hat{\lambda}_b - \lambda_b| |\hat{\lambda}_{\{R\}}|] (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}) + \frac{C'}{\sqrt{n}},$$

as soon as the event  $\mathcal{R}$  holds. The same reasoning and  $|\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}| \leq 2$  leads to

$$\begin{aligned} \mathbb{E} [T_{R_1, R_2} | \hat{\lambda}_{\{R\}}] &\leq C \sup_{S \subset [k] \setminus R, T \subset [k] \setminus R} \sum_{a \in S, b \in T} \frac{\sqrt{\lambda_b \lambda_a}}{n} (\mathbf{Q}_{ab} - \mathbf{Q}_{ab}^{ad,+}) + C'' \left( \frac{k}{n \sqrt{\log(k)}} + \frac{1}{\sqrt{n}} \right) \\ &\leq C \sqrt{\frac{k}{n \log(k)}}, \end{aligned}$$

as soon as the event  $\mathcal{R}$  holds. Going back to (71) and integrating the deviation inequality with respect to  $\hat{\lambda}_{\{R\}}$ , we conclude that

$$\mathbb{P} \left[ T_{R_1, R_2} \geq C \sqrt{\frac{k}{n \log(k)}} + 8 \sqrt{\frac{2t}{n}} \right] \geq 1 - e^{-t} - \mathbb{P}[\overline{\mathcal{R}}] \geq 1 - e^{-t} - 2e^{-\sqrt{k}/\log(k)}$$

where we use  $\mathbb{P}(\mathcal{R}) \geq 1 - 2e^{-\sqrt{k}/\log(k)}$ . From this point the proof is identical to that of the main proof: we fix  $t = 2 \log(k)q + \sqrt{k}/\log(k)$  and take an union bound over all possible  $R_1$  and  $R_2$  to derive that

$$\max_{R_1, R_2: |R_1| \leq q, |R_2| \leq q} Z_{R_1, R_2} \leq C \sqrt{\frac{k}{n \log(k)}} + 4q \max_{a=1, \dots, k} |\hat{\lambda}_a - \lambda_a|$$

on an event of probability higher than  $1 - 3 \exp(-\sqrt{k}/\log(k))$ . Then, as in the main proof, Lemma 6 together with (54) and (67) enforce that  $\|V\|_{\square}^{\pm} \leq C \sqrt{K/(n \log(k))}$  with probability larger than  $1 - (5 + 2k) \exp(-\sqrt{k}/\log(k))$ . By symmetry, we can find an event  $\mathcal{A}$  of probability larger than  $1 - (10 + 4k) \exp(-\sqrt{k}/\log(k))$  such that, on  $\mathcal{A}$ ,

$$\|V\|_{\square} \leq C \sqrt{\frac{k}{n \log(k)}}.$$

In order to control  $\mathbb{E}[\|V\|_{\square}]$  on the complementary event  $\bar{\mathcal{A}}$  we use the rough bound

$$\|V\|_{\square} \leq \|V\|_1 \leq \sum_{a, b=1}^k |\hat{\lambda}_a - \lambda_a| |\hat{\lambda}_b - \lambda_b| \leq 2 \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a|$$

which implies

$$\begin{aligned} \mathbb{E}[\|V\|_{\square}] &\leq \mathbb{E}[\|V\|_{\square} \mathbb{1}_{\mathcal{A}}] + \mathbb{E}[\|V\|_{\square} \mathbb{1}_{\bar{\mathcal{A}}}] \\ &\leq C \sqrt{\frac{k}{n \log(k)}} + 2 \mathbb{P}^{1/2}[\bar{\mathcal{A}}] \left[ \mathbb{E} \left( \sum_{a=1}^k |\hat{\lambda}_a - \lambda_a| \right)^2 \right]^{1/2} \\ &\leq C \sqrt{\frac{k}{n \log(k)}} + C' e^{-\sqrt{k}/(2 \log(k))} \frac{k}{\sqrt{n}} \leq C'' \sqrt{\frac{k}{n \log(k)}} \end{aligned}$$



where we use (52). Together with the decomposition (68), (69) and (70), we conclude that

$$\mathbb{E} \left[ \left\| (W - \widehat{W})|_{\mathcal{R} \times \mathcal{R}} \right\|_{\square} \right] \leq C \sqrt{\frac{k}{n \log(k)}}.$$

□

## F Proof of Theorem 2

It is enough to prove separately the following two minimax lower bounds:

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_{\square}(\widehat{f}, \rho_n W_0)] \geq C \rho_n \sqrt{\frac{k}{n \log(k)}}, \quad (78)$$

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[2]} \mathbb{E}_{W_0}[\delta_{\square}(\widehat{f}, \rho_n W_0)] \geq C \left( \sqrt{\frac{\rho_n}{n}} \wedge \rho_n \right). \quad (79)$$

The proof of (79) is identical to the proof of (45) in [26] so we just sketch the main idea. Fix some  $0 < \epsilon \leq 1/4$ . We consider  $W_1$  to be the constant graphon with  $W_1(x, y) \equiv 1/2$ , and  $W_2 \in \mathcal{W}^+[2]$  to be the 2-step graphon with  $W_2(x, y) = 1/2 + \epsilon$  if  $x, y \in [0, 1/2]^2 \cup [1/2, 1]^2$  and  $W_2(x, y) = 1/2 - \epsilon$  elsewhere. Obviously, we have  $\delta_{\square}[\rho_n W_1, \rho_n W_2] = \rho_n \epsilon$ . Then, standard testing arguments [32] ensure that the minimax risk  $\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[2]} \mathbb{E}_{W_0}[\delta_{\square}(\widehat{f}, \rho_n W_0)]$  is at least of the order  $\rho_n \epsilon$  when  $\epsilon$  is chosen small enough so that the  $\chi^2$ -distance  $\chi^2(\mathbb{P}_{W_2}, \mathbb{P}_{W_1})$  is smaller than  $1/4$ . According to Lemma 4.9 in [26], this is the case when  $\epsilon$  is small in front of  $(\rho_n n)^{-1/2}$  which proves (79).

Henceforth, we only focus on (78). We first consider the case of  $k$  multiple of 32 and such that  $k \geq C_0$  and  $k \leq C_1 n$  for some sufficiently large numerical constants  $C_0$  and  $C_1$ . As the collections  $\mathcal{W}^+[k]$  are nested this will imply (78) for all  $k \in [32C_0, n]$ . Afterwards, it will suffice to show (78) for  $k = 2$  to prove the proposition. So, we assume that  $k$  is a multiple of 32,  $k$  is large enough and that  $k$  is small in front of  $n$ . Define  $k_1 := k/2$ ,  $M_k := \lceil 128 \log(k) \rceil$ ,  $\eta_0 := 1/16$  and  $\eta_1 := 7/8$ .

As for Proposition 3, we will rely on Fano's method (Lemma 2). Hence, we shall build a collection  $(W_u)$  of graphons that are well-spaced in cut distance and such that the Kullback-Leibler divergence between the associated distribution  $\mathbb{P}_{W_u}$  remains small enough. All the graphons considered in this collection will be based on a  $k_1 \times M_k$  matrix  $\mathbf{B}$  such that (i) the rows of  $\mathbf{B}$  are almost orthogonal and (ii) such that the  $l_1$  distance between permutation and convex combinations of the columns of  $\mathbf{B}$  are bounded from below.

**Lemma 11.** *For  $k$  large enough, there exists a matrix  $\mathbf{B} \in \{-1, 1\}^{k_1 \times M_k}$  satisfying the following two properties:*

(i) *For any  $(a, b) \in [k_1]$  with  $a \neq b$ , the inner product of two columns  $\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle$  satisfies*

$$|\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle| \leq M_k/4. \quad (80)$$

(ii) *For any two subsets  $X$  and  $Y$  of  $[k_1]$  satisfying  $|X| = |Y| = \eta_0 k_1$  and  $X \cap Y = \emptyset$ , any labellings  $\pi_1 : [\eta_0 k_1] \rightarrow X$  and  $\pi_2 : [\eta_0 k_1] \rightarrow Y$ , any subset  $Z$  of  $[M_k]$  of size larger than  $\eta_1 M_k$  and any  $Z \times M_k$  stochastic matrix  $\omega$ , we have*

$$\sum_{a=1}^{\eta_0 k_1} \sum_{b \in Z} |\mathbf{B}_{\pi_1(a), b} - \sum_{c \in [M_k]} \omega_{b,c} \mathbf{B}_{\pi_2(a), c}| \geq C M_k k_1, \quad (81)$$

for some universal constant  $C > 0$ .

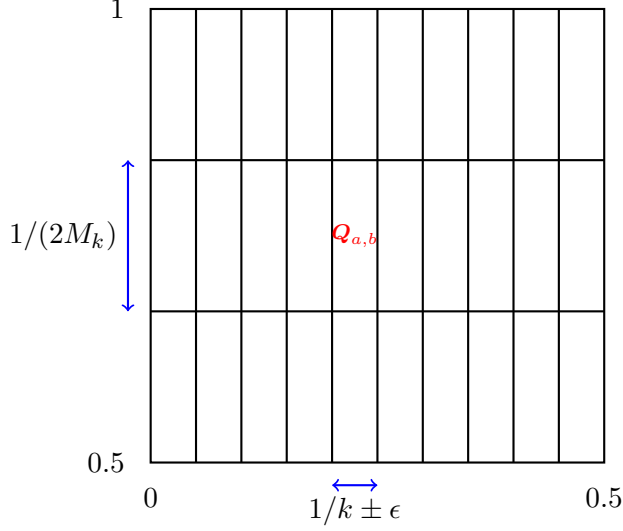


Figure 1: Restriction of  $W_u$  to  $[0, 1/2] \times [1/2, 1]$ .

Taking  $\mathbf{B}$  as in Lemma 11, we define the connection probability matrix  $\mathbf{Q} := (\mathbf{J} + \mathbf{B})/2$  where  $\mathbf{J}$  is the  $k_1 \times M_k$  matrix with all entries equal to 1.

Fix some  $\epsilon < 1/(8k_1)$  and denote by  $\mathcal{C}_0$  the collection of vectors  $u \in \{-\epsilon, \epsilon\}^{k_1}$  satisfying  $\sum_{a=1}^{k_1} u_a = 0$ . For any  $u \in \mathcal{C}_0$ , define the cumulative distribution  $F_u$  on  $\{0, \dots, k_1\}$  by the relations  $F_u(0) = 0$  and  $F_u(a) = a/(2k_1) + \sum_{b=1}^a u_b$  for  $a \in [k_1]$  and the cumulative distribution  $G$  on  $\{0, \dots, M_k\}$  by  $G(0) = 1/2$  and  $G(b) = 1/2 + b/(2M_k)$ . Note that  $F_u$  take values in  $[0, 1/2]$  and  $G$  takes values in  $[1/2, 1]$ . Then, set  $\Pi_{ab}(u) = [F_u(a-1), F_u(a)) \times [G(b-1), G(b))$  and define the graphon  $W_u \in \mathcal{W}^+[k_1 + M_k]$  by

$$W_u(x, y) = \begin{cases} \sum_{(a,b) \in [k_1] \times [M_k]} \mathbf{Q}_{ab} \mathbf{1}_{\Pi_{ab}(u)}(x, y) & \text{if } x \in [0, 1/2] \text{ and } y \in (1/2, 1] \\ W_u(y, x) & \text{if } x \in (1/2, 1] \text{ and } y \in [0, 1/2] \\ 1/2 & \text{else .} \end{cases}$$

See Figure (1) for a drawing of  $W_u$ . Note that  $W_u$  is a fairly unbalanced  $(k_1 + M_k)$ -step graphon:  $M_k$  of its steps have a large weight of order  $1/\log(k)$ . Besides, the  $k_1$  smaller steps are slightly unbalanced as the weight of each class is either  $1/k - \epsilon$  or  $1/k + \epsilon$ . The purpose of these  $M_k$  big steps is to make the cut distances between  $W_u$  and  $W_v$  the largest possible.

Next, we shall consider a subcollection  $\mathcal{C}$  of  $\mathcal{C}_0$  such that the graphons  $W_u$  with  $u \in \mathcal{C}$  are well spaced. The following combinatorial result is in the spirit of the Varshamov-Gilbert lemma [32, Lemma 2.9]. It is borrowed from [26] (Lemma 4.4). For  $u \in \mathcal{C}_0$ , let  $\mathcal{A}_u := \{a \in [k_1] : u_a = \epsilon\}$ . Notice that, by definition of  $\mathcal{C}_0$ , we have  $|\mathcal{A}_u| = k_1/2$  for all  $u \in \mathcal{C}_0$ .

**Lemma 12.** *There exists a subset  $\mathcal{C}$  of  $\mathcal{C}_0$  such that  $\log |\mathcal{C}| \geq k_1/16$  and*

$$|\mathcal{A}_u \Delta \mathcal{A}_v| > k_1/4 . \quad (82)$$

for any  $u \neq v \in \mathcal{C}$ .

Lemmas 11 and 12 are used to obtain the following lower bound on the distance  $\delta_{\square}(W_u, W_v)$  between two distinct graphons with  $u$  and  $v$  in  $\mathcal{C}$ . This lemma is the main ingredient of the proof.

**Lemma 13.** *There exists two positive universal constants  $C_1$  and  $C_2$  such that if  $k\epsilon \leq C_2$  then, for any  $(u, v) \in \mathcal{C}$  with  $u \neq v$ , we have*

$$\delta_{\square}(W_u, W_v) \geq C_1 \frac{k\epsilon}{\sqrt{M_k}} \quad (83)$$

which implies

$$\delta_{\square}(\rho_n W_u, \rho_n W_v) \geq C_1 \rho_n \frac{k\epsilon}{\sqrt{M_k}}. \quad (84)$$

Note that for any  $u$  and  $v$  in  $\mathcal{C}$  it is possible to build a measure-preserving transformation  $\tau$  such that  $W_u - W_v^\tau$  is null except on a measurable set of Lebesgue measure of order  $k\epsilon$  (see the proof of Theorem 1 in Section E for such construction). Hence, the  $l_1$  norm of  $W_u - W_v^\tau$  is of order  $k\epsilon$ . Lemma 13 states, that by taking the infimum over all  $\tau$  and by considering the weaker norm  $\|\cdot\|_{\square}$ , one still has a lower bound of the same order. The  $M_k^{-1/2}$  factor arises as a consequence of Lemma 4. See the proof for more details.

To apply Fano's method, we need to upper bound the Kullback-Leibler divergence between the distribution corresponding to any two graphon  $W_u$  and  $W_v$  with  $u$  and  $v$  in  $\mathcal{C}$ . Let  $\mathbb{P}_{W_u}$  denote the distribution of  $\mathbf{A}$  sampled according to the sparse  $W$ -random graph model (1) with  $W_0 = W_u$ . Since the matrix  $\mathbf{Q}$  is fixed the difficulty in distinguishing between the distributions  $\mathbb{P}_{W_u}$  and  $\mathbb{P}_{W_v}$  for  $u \neq v$  comes from the randomness of the design points  $\xi_1, \dots, \xi_n$  in the  $W$ -random graph model (1) rather than from the randomness of the realization of the adjacency matrix  $\mathbf{A}$  conditionally on  $\xi_1, \dots, \xi_n$ . The following lemma gives an upper bound on the Kullback-Leibler divergences  $\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v})$ :

**Lemma 14.** *For all  $u, v \in \mathcal{C}_0$  we have*

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq 32nk_1^2\epsilon^2/3.$$

Now, choose  $\epsilon$  such that  $\epsilon^2 = \frac{3}{(16)^{3nk_1}}$ . When  $k$  is small in front of  $n$ , this choice of  $\epsilon$  satisfies the conditions of Lemma 13. Then it follows from Lemmas 12 and 14 that

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq \frac{1}{16} \log |\mathcal{C}|, \quad \forall u, v \in \mathcal{C} : u \neq v. \quad (85)$$

In view Fano's Lemma (Lemma 2), inequalities (84) and (85) imply that

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_{\square}(\hat{f}, \rho_n W_0)] \geq C \rho_n \sqrt{\frac{k}{n \log(k)}}$$

where  $C > 0$  is an absolute constant. This completes the proof for  $k$  large enough.

Now we turn to the case  $k = 2$ . We reduce the lower bound to the problem of testing two hypotheses. Consider the matrix  $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ . Given  $u \in \{-\epsilon, +\epsilon\}$  define  $F_u(0) = 0$ ,  $F_u(1) = 1/2 + u$  and  $F_u(2) = 1$ . Then, we set  $\Pi_{ab}(u) = [F_u(a-1), F_u(a)) \times [F_u(b-1), F_u(b))$  for any  $a, b \in \{1, 2\}$  and define graphons

$$W_u(x, y) := \sum_{a,b=1}^2 \frac{(\mathbf{B}_{ab} + 1)}{2} \mathbb{1}_{\Pi_{ab}(u)}(x, y).$$

For any measure preserving bijection  $\tau$ ,  $(W_\epsilon - W_{-\epsilon}^\tau)$  is a four-step graphon. Thanks to Lemma 4, we deduce that  $\delta_\square(W_\epsilon, W_{-\epsilon}) \geq C\delta_1(W_\epsilon, W_{-\epsilon})$ . Then, it is not hard to see that  $\delta_1(W_\epsilon, W_{-\epsilon}) \geq C'\epsilon$  so that  $\delta_\square(\rho_n W_\epsilon, \rho_n W_{-\epsilon}) \geq C'\rho_n\epsilon$ . Moreover, exactly as in Lemma 14, the Kullback-Leibler divergence between  $\mathbb{P}_{W_\epsilon}$  and  $\mathbb{P}_{W_{-\epsilon}}$  is bounded by  $Cn\epsilon^2$ . Taking  $\epsilon$  of the order  $n^{-1/2}$ , this divergence is small. It is therefore impossible to reliably distinguish  $\mathbb{P}_{W_\epsilon}$  from  $\mathbb{P}_{W_{-\epsilon}}$  and the estimation error is at least of order  $\rho_n\epsilon$ . More formally, we use Theorem 2.2 from [32] to conclude that

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[2]} \mathbb{E}_{W_0}[\delta_\square(\hat{f}, \rho_n W_0)] \geq C\rho_n \sqrt{\frac{1}{n}}$$

where  $C > 0$  is an absolute constant.

**Proof of Lemma 11.** Let  $\mathbf{B}$  be a  $k_1 \times M_k$  random matrix whose entries are independent Rademacher variables. We shall prove that, with positive probability,  $\mathbf{B}$  satisfies both (80) and (81).

Fix  $a \neq b$ . Then,  $\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle$  is distributed as a sum of  $k_1$  independent Rademacher variables. Using Hoeffding's inequality, we have that

$$\mathbb{P}[|\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle| \geq M_k/4] \leq 2\exp[-M_k/32].$$

By the union bound, property (80) is satisfied for all  $a \neq b$  with probability greater than  $1 - k_1^2 \exp[-M_k/32]$ . Since  $M_k \geq 128 \log(k)$ , for  $k$  greater than some absolute constant, this probability is greater than 3/4.

Turning to (81), we first fix  $X, Y, Z, \pi_1, \pi_2$ , and  $\omega$ . Let

$$T_{X,Y,Z,\pi_1,\pi_2,\omega} := \sum_{a=1}^{\eta_0 k_1} \sum_{b \in Z} |\mathbf{B}_{\pi_1(a),b} - \sum_{c \in [M_k]} \omega_{b,c} \mathbf{B}_{\pi_2(a),c}|.$$

We have that, conditionally on  $(\mathbf{B}_{b,c})_{b \in Y, c \in [M_k]}$ ,  $T_{X,Y,Z,\pi_1,\pi_2,\omega}$  stochastically dominates a binomial distribution with parameters  $(\eta_0 k_1) \times |Z|$  and 1/2. Then, Hoeffding's inequality yields

$$\mathbb{P}\{T_{X,Y,Z,\pi_1,\pi_2,\omega} \leq \eta_0 k_1 |Z|/4\} \leq 2\exp(-\eta_0 \eta_1 k_1 M_k/8).$$

Given any integer  $Z \in [\eta_1 M_k, M_k]$ , define  $\Omega_Z$  the collection of  $Z \times [M_k]$  stochastic matrices taking values in the discrete set  $\{0, 1/(8M_k), 2/(8M_k), \dots, 1\}$ . Since  $X, Y \subset [k_1]$  and  $Z \subset M_k$ , it is easy to see that the cardinality of the set of all possible tuples  $(X, Y, Z, \pi_1, \pi_2, \omega)$  with  $\omega \in \Omega_Z$  is bounded by

$$2^{2k_1 + M_k} ((\eta_0 k_1)!)^2 (8M_k + 1)^{M_k^2}.$$

Now, taking the union bound, we derive that, simultaneously for all such parameters,

$$T_{X,Y,Z,\pi_1,\pi_2,\omega} > \eta_0 k_1 |Z|/4$$

with probability greater than  $1 - 2^{2k_1 + M_k + 1} (\eta_0 k_1)!^2 (8M_k + 1)^{M_k^2} \exp[-\eta_0 \eta_1 k_1 M_k/8]$ . Using Stirling's approximation

and  $\eta_1 M_k \geq 64 \log(k)$  we get that this probability is larger than 3/4 for  $k$  large enough.

Finally, let us consider a general case, when matrix  $\omega$  does not necessarily belong to  $\Omega_Z$ . Observe that in this case, there exists a matrix  $\omega' \in \Omega_Z$  such that  $\max_{b \in Z} \sum_{c \in [M_k]} |\omega_{b,c} - \omega'_{b,c}| \leq 1/8$ . This implies that

$$T_{X,Y,Z,\pi_1,\pi_2,\omega} \geq T_{X,Y,Z,\pi_1,\pi_2,\omega'} - \frac{|Y||Z|}{8} \geq \eta_0 \eta_1 k_1 M_k/8.$$

We have proved that (81) holds with probability larger than 3/4. As a consequence,  $\mathbf{B}$  satisfies both (80) and (81) with probability larger than 1/2.  $\square$

**Proof of Lemma 13.** We fix  $u$  and  $v$ , two different vectors in  $\mathcal{C}$ , and fix  $\tau$ , a measure-preserving bijection on  $[0, 1] \rightarrow [0, 1]$ . We shall prove that for  $k\epsilon$  small enough

$$\|W_u - W_v^\tau\|_\square \geq C \frac{k\epsilon}{\sqrt{M_k}}. \quad (86)$$

Since  $\delta_\square(W_u, W_v) = \inf_\tau \|W_u(\cdot, \cdot) - W_v(\tau, \tau)\|_\square$  both (83) and (84) straightforwardly follow from (86). We denote

$$\begin{aligned} \mathcal{B}_{11} &:= \tau^{-1}([0, 1/2]) \cap [0, 1/2], & \mathcal{B}_{12} &:= \tau^{-1}([0, 1/2]) \cap (1/2, 1], \\ \mathcal{B}_{21} &:= \tau^{-1}((1/2, 1]) \cap [0, 1/2], & \mathcal{B}_{22} &:= \tau^{-1}((1/2, 1]) \cap (1/2, 1]. \end{aligned} \quad (87)$$

Since  $\tau$  is measure-preserving, we have

$$\lambda(\mathcal{B}_{11}) = \lambda(\mathcal{B}_{22}) = 1/2 - \lambda(\mathcal{B}_{12}) = 1/2 - \lambda(\mathcal{B}_{21}). \quad (88)$$

Now, we consider three cases (i)  $\lambda(\mathcal{B}_{12}) \leq k_1\epsilon/64$ , (ii)  $k_1\epsilon/64 < \lambda(\mathcal{B}_{12}) \leq 1/2 - k_1\epsilon/64$  and (iii)  $\lambda(\mathcal{B}_{12}) > 1/2 - k_1\epsilon/64$ . In the Case (i) we shall focus on the restriction of  $W_u$  and  $W_v^\tau$  on  $\mathcal{B}_{11} \times \mathcal{B}_{22}$  so that these restrictions are  $k_1 \times M_k$ -step functions. In the Case (ii), we focus on restrictions to  $\mathcal{B}_{21} \times \mathcal{B}_{22}$ , so that  $W_v^\tau$  is constant on this restriction. In the pathological case (iii), we introduce a subset such that the restriction of  $W_u$  is a  $M_k \times k_1$ -step function and the restriction of  $W_v^\tau$  is a  $k_1 \times M_k$ -step function.

**Case (i).** We focus our attention on coordinates  $(x, y)$  in  $\mathcal{B}_{11} \times \mathcal{B}_{22}$ . Recall that the cumulative distribution function  $G$  is defined by  $G(0) = 1/2$  and  $G(b) = 1/2 + b/(2M_k)$  for  $b \in [M_k]$ . For any  $(r, s) \in [M_k]^2$ , define

$$\omega_{r,s} := \lambda\{[G(r-1), G(r)) \cap \tau^{-1}([G(s-1), G(s)))\}.$$

By definition of  $\omega_{r,s}$ , for any  $r \in [M_k]$ , we have

$$\omega_{r\bullet} := \sum_{s \in [M_k]} \omega_{r,s} \leq 1/(2M_k) \quad \text{and} \quad \sum_{r,s} \omega_{r,s} = \lambda(\mathcal{B}_{22}).$$

Let  $\mathcal{R}$  denote the sets of  $r \in [M_k]$  such that  $[G(r-1), G(r))$  has a large intersection with  $\tau^{-1}([1/2, 0])$ :

$$\mathcal{R} := \{r \in [M_k] \text{ s.t. } \omega_{r\bullet} \geq 3/(7M_k)\} \quad \text{and} \quad \mathcal{Y} := \cup_{r \in \mathcal{R}} [G(r-1), G(r)) \cap \mathcal{B}_{22}. \quad (89)$$

Denote  $\bar{\mathcal{R}}$  the complementary set of  $\mathcal{R}$ . We have that  $\lambda(\mathcal{B}_{22}) = 1/2 - \lambda(\mathcal{B}_{12}) \geq 1/2 - k_1\epsilon/64 \geq \frac{27}{56}$  for  $k_1\epsilon$  small enough. Hence, it follows that

$$\frac{27}{56} \leq \sum_{r,s} \omega_{r,s} = \sum_{r \in [M_k]} \omega_{r\bullet} = \sum_{r \in \mathcal{R}} \omega_{r\bullet} + \sum_{r \in \bar{\mathcal{R}}} \omega_{r\bullet} \quad (90)$$

which implies that  $|\mathcal{R}| \geq 3M_k/4$  and  $\lambda(\mathcal{Y}) = \sum_{r \in \mathcal{R}} \omega_{r\bullet} \geq 9/28$ .

Now, denoting  $\mathcal{X} := \mathcal{B}_{11}$ , we define a new kernel  $\bar{W}_v^\tau : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$  by

$$\begin{aligned} \bar{W}_v^\tau(x, y) &:= \sum_{r \in \mathcal{R}} \mathbb{1}_{\{y \in [G(r-1), G(r))\}} \frac{1}{\lambda\{[G(r-1), G(r)) \cap \mathcal{Y}\}} \int_{[G(r-1), G(r)) \cap \mathcal{Y}} W_v(\tau(x), \tau(z)) dz \\ &= \sum_{a=1}^{k_1} \sum_{r \in \mathcal{R}} \mathbb{1}_{\{y \in [G(r-1), G(r))\}} \mathbb{1}_{\{\tau(x) \in [F_v(a-1), F_v(a))\}} \sum_{s \in [M_k]} \frac{\omega_{r,s} (1 + \mathbf{B}_{as})}{\omega_{r\bullet} 2}. \end{aligned} \quad (91)$$

We can view  $\overline{W}_v^\tau$  as a smoothed version of the restriction of  $W_v^\tau$  to  $\mathcal{X} \times \mathcal{Y}$ . The marginal functions  $\overline{W}_v^\tau(x, \cdot)$  are step functions with at most  $|\mathcal{R}| \leq M_k$  steps of the form  $[G(r-1), G(r)) \cap \mathcal{B}_{22}$ . Moreover, on each interval  $[G(r-1), G(r)) \cap \mathcal{B}_{22}$ ,  $\overline{W}_v^\tau(x, y)$  is equal to the mean of  $W_v^\tau(x, z)$  for  $z$  ranging on this set. Equipped with this notation, we can control the cut distance between  $W_u$  and  $W_v^\tau$  in terms of the  $l_1$  distance between the restriction of  $W_u$  to  $\mathcal{X} \times \mathcal{Y}$  and  $\overline{W}_v^\tau$ . For ease of notation, we still write  $W_u$  for the restriction of  $W_u$  to  $\mathcal{X} \times \mathcal{Y}$  when there is no ambiguity.

The following lemma provides a lower bound of the cut norm  $\|W_u - W_v^\tau\|_\square$  in terms of the  $l_1$  norm of  $\|W_u - \overline{W}_v^\tau\|_1$ .

**Lemma 15.** *For any  $u, v$  in  $\mathcal{C}$  and any measure-preserving transformation  $\tau$ , we have*

$$\|W_u - W_v^\tau\|_\square \geq \frac{1}{4\sqrt{2M_k}} \|W_u - \overline{W}_v^\tau\|_1, \quad (92)$$

where  $\overline{W}_v^\tau$  is defined in (91).

In view of Lemma 15 it is enough to control the  $l_1$  norm  $\|W_u - \overline{W}_v^\tau\|_1$ . We can do it in a similar way as it is done in the proof of Lemma 4.5 in [26]. For  $a \neq b$  and any  $x \in [F_u(a-1), F_u(a)) \cap \mathcal{X}$  and  $x' \in [F_u(b-1), F_u(b)) \cap \mathcal{X}$ , the inner product between  $W_u(x, \cdot)$  and  $W_v(x', \cdot)$  satisfies

$$\begin{aligned} & \left| \int_{\mathcal{Y}} (W_u(x, y) - 1/2)(W_v(x', y) - 1/2) dy \right| \\ & \leq \left| \int_{[1/2, 1]} (W_u(x, y) - 1/2)(W_v(x', y) - 1/2) dy \right| + \frac{1}{4} \lambda\{[1/2, 1] \setminus \mathcal{Y}\} \\ & \leq \frac{1}{8M_k} \left| \sum_{c=1}^{M_k} B_{ac} B_{bc} \right| + \frac{5}{112} \leq \frac{1}{32} + \frac{5}{112} \end{aligned} \quad (93)$$

where we used (80) in the last line. For any  $a, b \in [k_1]$ , let  $\psi_{ab}$  denote the Lebesgue measure of the set

$$[F_u(a-1), F_u(a)) \cap \tau^{-1}([F_v(b-1), F_v(b))) \cap \mathcal{X}.$$

Since  $\tau$  is measure preserving, it follows that  $\sum_b \psi_{ab} \leq 1/(2k_1) + u_a$  and  $\sum_a \psi_{ab} \leq 1/(2k_1) + v_b$ . For any  $y \in \mathcal{Y}$ , we set

$$h_{u,a}(y) := W_u(F_u(a-1), y) - 1/2 \quad \text{and} \quad k_{v,b}(y) := \overline{W}_v^\tau(\tau^{-1}(F_v(b-1)), y) - 1/2.$$

Equipped with this notation, we have

$$\int_{\mathcal{X} \times \mathcal{Y}} |W_u(x, y) - \overline{W}_v^\tau(x, y)| dx dy = \sum_{a=1}^{k_1} \sum_{b=1}^{k_1} \psi_{a,b} \int_{\mathcal{Y}} |h_{u,a}(y) - k_{v,b}(y)| dy.$$

Now take any  $a_1 \neq a_2$ . By (93),  $|h_{u,a}(y)| = 1/2$  and using the triangle inequality, we derive that

$$\begin{aligned} \|h_{u,a_1} - k_{v,b}\|_1 + \|h_{u,a_2} - k_{v,b}\|_1 & \geq \|h_{u,a_1} - h_{u,a_2}\|_1 \\ & \geq \|h_{u,a_1} - h_{u,a_2}\|_2^2 \\ & \geq 2 \left[ \frac{1}{4} \lambda(\mathcal{Y}) - \frac{1}{32} - \frac{5}{112} \right] \geq \frac{1}{112}, \end{aligned}$$

where we used  $\lambda(\mathcal{Y}) \geq 9/28$  in the last line. As a consequence, for any  $b \in [k_1]$  there exists at most one  $a \in [k_1]$  such that  $\|h_{u,a} - k_{v,b}\|_1 < 1/224$ . If such index  $a$  exists, it is denoted by  $\pi(b)$ . Then, it is possible to extend  $\pi$  as a function from  $[k_1]$  to  $[k_1]$ . Since  $\sum_{a,b} \psi_{a,b} = \lambda(\mathcal{X})$ , we get

$$\begin{aligned} \|W_u - \overline{W}_v^\tau\|_1 &\geq \frac{1}{224} \sum_{b=1}^{k_1} \sum_{a \neq \pi(b)} \psi_{a,b} = \frac{1}{224} \sum_{b=1}^{k_1} \left[ (1/(2k_1) + v_b - \psi_{\pi(b),b}) - \left( \frac{1}{2} - \lambda[\mathcal{X}] \right) \right] \\ &= \frac{1}{224} \sum_{b=1}^{k_1} [(1/(2k_1) + v_b - \psi_{\pi(b),b}) - \lambda[\mathcal{B}_{1,2}]] \\ &\geq \frac{1}{224} \sum_{b=1}^{k_1} [(1/(2k_1) + v_b - \psi_{\pi(b),b}) - k_1\epsilon/64] , \end{aligned}$$

since  $\lambda[\mathcal{B}_{1,2}] \leq k_1\epsilon/64$ . If the sum  $\sum_{b=1}^{k_1} 1/(2k_1) + v_b - \psi_{\pi(b),b}$  is greater than  $k_1\epsilon/32$ , then (86) is satisfied. Thus, we can assume in the sequel that  $\sum_{b=1}^{k_1} 1/(2k_1) + v_b - \psi_{\pi(b),b} \leq k_1\epsilon/32$ .

Using that  $\psi_{a,b} \leq (1/(2k_1) + u_a) \wedge (1/(2k_1) + v_b)$  and that the cardinality of the collection  $\{b \in [k_1] : v_b > 0\}$  is  $k_1/2$  we deduce that the collection  $\{b \in [k_1] : v_b > 0, u_{\pi(b)} > 0 \text{ and } \psi_{\pi(b),b} \geq 1/(2k_1)\}$  has cardinality greater than  $7k_1/16$ . Now, Lemma 12 implies that  $|\mathcal{A}_u \cap \mathcal{A}_v| \leq 3k_1/8$  for  $u \neq v \in \mathcal{C}$ . Then, there exist subsets  $A \subset \mathcal{A}_u$  and  $B \subset \mathcal{A}_v$  of cardinality  $\eta_0 k_1$  (recall that  $\eta_0 = 1/16$ ) such that  $\pi(B) = A$ ,  $A \cap B = \emptyset$ , and  $\psi_{\pi(b),b} \geq 1/(2k_1)$  for all  $b \in B$ . The condition  $\psi_{\pi(b),b} \geq 1/(2k_1)$  implies that  $\pi$  is injective on  $B$ . Hence,

$$\begin{aligned} \|W_u - \overline{W}_v^\tau\|_1 &\geq \sum_{b \in B} \psi_{\pi(b),b} \int_{\mathcal{Y}} |h_{u,\pi(b)}(y) - k_{v,b}(y)| dy \\ &\geq \frac{C}{k_1 M_k} \sum_{b \in B} \sum_{c \in \mathcal{R}} \left| Q_{\pi(b),c} - \frac{\sum_{d \in [M_k]} \omega_{b,d} Q_{b,d}}{\omega_{b,\bullet}} \right| \\ &= \frac{C'}{k_1 M_k} \sum_{b \in B} \sum_{c \in \mathcal{R}} \left| B_{\pi(b),c} - \frac{\sum_{d \in [M_k]} \omega_{c,d} B_{b,d}}{\omega_{c,\bullet}} \right| , \end{aligned}$$

where the second inequality follows from  $\psi_{\pi(b),b} \geq 1/(2k_1)$  and the fact that  $h_{u,\pi(b)}$  and  $k_{v,b}$  are step functions with steps larger than  $3/(7M_k)$  (see (89), the definition of  $\mathcal{R}$  and  $\mathcal{Y}$ ). Finally, we apply the property (81) of  $B$  to conclude that

$$\int |W_u(x, y) - \overline{W}_v^\tau(x, y)| dx dy \geq C \geq C' k_1 \epsilon ,$$

which, together with Lemma 15, proves (86).

**Case (ii).** Now we assume that  $k_1\epsilon/64 < \lambda(\mathcal{B}_{12}) < 1/2 - k_1\epsilon/64$ . Take  $\mathcal{X} = \mathcal{B}_{21}$  and  $\mathcal{Y} = \mathcal{B}_{22}$ . We have that, on  $\mathcal{X} \times \mathcal{Y}$ ,  $W_v^\tau$  is constant and equals  $1/2$ . Denote  $U$  the restriction of  $W_u - 1/2$  to  $\mathcal{X} \times \mathcal{Y}$ . Then, it follows that  $\|W_u - W_v^\tau\|_\square \geq \|U\|_\square$ . The kernel  $U$  is at most  $k_1 \times M_k$  step function. By Lemma 4, we obtain

$$\|U\|_\square \geq \frac{1}{4\sqrt{2M_k}} \|U\|_1 = \frac{1}{8\sqrt{2M_k}} \lambda(\mathcal{X}) \lambda(\mathcal{Y}) = \frac{1}{8\sqrt{2M_k}} \lambda(\mathcal{X}) \left( \frac{1}{2} - \lambda(\mathcal{X}) \right) ,$$

where the last equality follows from (88). Using  $\lambda(\mathcal{X}) = \lambda(\mathcal{B}_{12})$  and  $x(1/2-x) \geq 1/4 \min(x, (1/2-x))$  we obtain (86).



**Case (iii).** Now we assume that  $\lambda(\mathcal{B}_{12}) \geq 1/2 - k\epsilon/64$  and take  $\mathcal{X} = \mathcal{B}_{21}$  and  $\mathcal{Y} = \mathcal{B}_{12}$  so that  $\lambda(\mathcal{X}) = \lambda(\mathcal{B}_{12}) \geq 1/2 - k_1\epsilon/64$ . Define the smoothed kernel  $\overline{W}_v^\tau : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$  by

$$\overline{W}_v^\tau(x, y) := \sum_{a=1}^{M_r} \mathbb{1}_{\{y \in [G(a-1), G(a))\}} \frac{1}{\lambda\{[G(a-1), G(a)) \cap \mathcal{Y}\}} \int_{[G(a-1), G(a)) \cap \mathcal{Y}} W_v(\tau(x), \tau(z)) dz.$$

As a consequence,  $\overline{W}_v^\tau$  is  $M_k \times M_k$  block-constant on subsets of the form  $(\tau^{-1}[G(a-1), G(a)) \cap \mathcal{X}) \times ([G(b-1), G(b)) \cap \mathcal{Y})$ . Arguing as in the proof of Lemma 15, we derive that

$$\|W_u - W_v^\tau\|_\square \geq \frac{1}{4\sqrt{2M_k}} \|W_u - \overline{W}_v^\tau\|_1. \quad (94)$$

For any  $a$  such that  $[F_u(a-1), F_u(a)) \cap \mathcal{X} \neq \emptyset$  define the function  $h_{u,a}$  on  $\mathcal{Y}$  by  $h_{u,a}(y) := W_u(F_u(a-1), y) - 1/2$ . Arguing as in Case (i), we observe that  $\|h_{u,a_1} - h_{u,a_2}\|_1 \geq 1/112$  for any  $a_1 \neq a_2$ . We have that the kernel  $\overline{W}_v^\tau$  is a  $M_k \times M_k$  step function. Hence, there exists a partition  $(\mathcal{X}_b)_{b=1, \dots, M_k}$  of  $\mathcal{X}$  and  $M_k$  functions  $k_b(y)$  such that  $(\overline{W}_v^\tau - 1/2)(x, y) = \sum_{b=1}^{M_k} \mathbb{1}_{x \in \mathcal{X}_b} k_b(y)$ . Then, the triangular inequality ensures that, for any  $a_1 \neq a_2$  and any  $b \in [M_k]$ , we have  $\|h_{u,a_1} - k_b\|_1 + \|h_{u,a_1} - k_b\|_1 \geq \|h_{u,a_1} - h_{u,a_2}\|_1 \geq 1/112$ . As a consequence, for any  $b \in [M_k]$  there exists at most one  $a$ , which we will denote by  $\pi(b)$ , such that  $\|h_{u,\pi(b)} - k_b\|_1 \leq 1/224$ . Now we compute

$$\begin{aligned} \|W_u - \overline{W}_v^\tau\|_1 &= \sum_{b=1}^{M_k} \sum_{a=1}^{k_1} \lambda(\mathcal{X}_b \cap [F_u(a-1), F_u(a)) \cap \mathcal{X}) \|h_{u,a} - k_b\|_1 \\ &\geq \frac{1}{224} \sum_{b=1}^{M_k} \lambda[\mathcal{X}_b \setminus [F_u(\pi(b)-1), F_u(\pi(b))) \cap \mathcal{X}] \\ &\geq \frac{1}{224} \left[ \lambda(\mathcal{X}) - \sum_{b=1}^{M_k} \frac{1}{2k_1} + u_{\pi(b)} \right] \\ &\geq \frac{1}{224} \left[ \lambda(\mathcal{X}) - \frac{M_k}{2k_1} - M_k\epsilon \right] \geq C', \end{aligned}$$

where we used  $\lambda(\mathcal{X}) \geq 1/4$ ,  $M_k/k \leq 1/8$ , and that  $M_k\epsilon \leq k\epsilon$  is small enough. Together with (94), we obtain the desired result (86).  $\square$

**Proof of Lemma 15.** We first prove that  $\|W_u - \overline{W}_v^\tau\|_\square \leq \|W_u - W_v^\tau\|_\square$ . Fix any measurable subset  $S \subset \mathcal{X}$ . Since functions  $[W_u - \overline{W}_v^\tau](x, \cdot)$  are constant on each set  $[G(r-1), G(r)) \cap \mathcal{Y}$ , the supremum  $\sup_{T \subset \mathcal{Y}} \left| \int_{S \times T} W_u(x, y) - \overline{W}_v^\tau(x, y) dx dy \right|$  is achieved by a subset  $T$  which is an union of some of  $[G(r-1), G(r)) \cap \mathcal{Y}$ , that is  $T = \cup_{r \in \mathcal{R}' \subset \mathcal{R}} [G(r-1), G(r)) \cap \mathcal{Y}$ . For such  $T$ , the definition (91) of  $\overline{W}_v^\tau$  implies  $\int_{S \times T} \overline{W}_v^\tau(x, y) dx dy = \int_{S \times T} W_v^\tau(x, y) dx dy$  so that

$$\sup_{T \subset \mathcal{Y}} \left| \int_{S \times T} W_u(x, y) - \overline{W}_v^\tau(x, y) dx dy \right| \leq \sup_{T \subset \mathcal{Y}} \left| \int_{S \times T} W_u(x, y) - W_v^\tau(x, y) dx dy \right|.$$

Taking the supremum over all  $S$  leads to  $\|W_u - \overline{W}_v^\tau\|_\square \leq \|W_u - W_v^\tau\|_\square$ . By definition of  $W_u$  and  $\overline{W}_v^\tau$  we have that  $U$  is a  $k_1^2 \times M_k$  step function. Then, Lemma 4 allows us to conclude

$$\|W_u - W_v^\tau\|_\square \geq \frac{1}{4\sqrt{2M_k}} \|W_u - \overline{W}_v^\tau\|_1.$$

$\square$

**Proof of Lemma 14.** The proof of Lemma 14 follows the lines of the proof of Lemma 4.3 in [26] and we give it here for completeness. For  $u \in \mathcal{C}_0$ , let  $\zeta(u) = (\zeta_1(u), \dots, \zeta_n(u))$  be the vector of  $n$  i.i.d. random variables with the discrete distribution on  $[k_1 + M_k]$  defined by

$$\mathbb{P}[\zeta_1(u) = a] = \begin{cases} 1/(2k_1) + u_a & \text{if } a \in [k_1] \\ 1/(2M_k) & \text{if } k_1 + 1 \leq a \leq M_k + k_1 \end{cases} \quad (95)$$

Let  $\Theta_0$  be the  $n \times n$  symmetric matrix with elements  $(\Theta_0)_{ii} = 0$  and  $(\Theta_0)_{ij} = \rho_n \mathbf{Q}_{\zeta_i(u), \zeta_j(u)}$  for  $i \neq j$ . Assume that, conditionally on  $\zeta(u)$ , the adjacency matrix  $\mathbf{A}$  is sampled according to the network sequence model with such probability matrix  $\Theta_0$ . Notice that in this case the observations  $\mathbf{A}' = (\mathbf{A}_{ij}, 1 \leq j < i \leq n)$  have the probability distribution  $\mathbb{P}_{W_u}$ . Using this remark and introducing the probabilities  $\alpha_{\mathbf{a}}(u) = \mathbb{P}[\zeta(u) = \mathbf{a}]$  and  $p_{A\mathbf{a}} = \mathbb{P}[\mathbf{A}' = A | \zeta(u) = \mathbf{a}]$  for  $\mathbf{a} \in [k_1 + M_k]^n$ , we can write the Kullback-Leibler divergence between  $\mathbb{P}_{W_u}$  and  $\mathbb{P}_{W_v}$  in the form

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) = \sum_A \sum_{\mathbf{a}} p_{A\mathbf{a}} \alpha_{\mathbf{a}}(u) \log \left( \frac{\sum_{\mathbf{a}} p_{A\mathbf{a}} \alpha_{\mathbf{a}}(u)}{\sum_{\mathbf{a}} p_{A\mathbf{a}} \alpha_{\mathbf{a}}(v)} \right)$$

where the sums in  $\mathbf{a}$  are over  $[k_1 + M_k]^n$  and the sum in  $A$  is over all triangular upper halves of matrices in  $\{0, 1\}^{n \times n}$ . Since the function  $(x, y) \mapsto x \log(x/y)$  is convex we can apply Jensen's inequality to get

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq \sum_{\mathbf{a}} \alpha_{\mathbf{a}}(u) \log \left( \frac{\alpha_{\mathbf{a}}(u)}{\alpha_{\mathbf{a}}(v)} \right) = n \sum_{a \in [k_1 + M_k]} \mathbb{P}[\zeta_1(u) = a] \log \left( \frac{\mathbb{P}[\zeta_1(u) = a]}{\mathbb{P}[\zeta_1(v) = a]} \right)$$

where the last equality follows from the fact that  $\alpha_{\mathbf{a}}(u)$  are  $n$ -product probabilities. Using (95) we get

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq n \sum_{a \in [k_1]} (1/(2k_1) + u_a) \log \left( \frac{1/(2k_1) + u_a}{1/(2k_1) + v_a} \right), \quad (96)$$

which is equal to  $n/2$  times the Kullback-Leibler divergence between two discrete distribution. Since the Kullback-Leibler divergence is less than the chi-square divergence we obtain

$$\sum_{a \in [k_1]} (1/k_1 + 2u_a) \log \left( \frac{1/k_1 + 2u_a}{1/k_1 + 2v_a} \right) \leq \sum_{a \in [k_1]} \frac{4(u_a - v_a)^2}{1/k_1 + 2v_a} \leq 64k^2 \epsilon^2 / 3,$$

where last inequality we use  $|v_a| \leq \epsilon \leq 1/(8k_1)$ , and  $|u_a - v_a| \leq 2\epsilon$ . Combining this with (96) proves the lemma.  $\square$

## G Proof of Proposition 6

To prove (23), it is enough to prove separately the following three minimax lower bounds:

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_1(\hat{f}, \rho_n W_0)] \geq C \rho_n \sqrt{\frac{k-1}{n}}, \quad (97)$$

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_1(\hat{f}, \rho_n W_0)] \geq C \min \left( \sqrt{\rho_n \frac{k}{n}}, \rho_n \right), \quad (98)$$

$$\inf_{\hat{f}} \sup_{W_0 \in \mathcal{W}^+[2]} \mathbb{E}_{W_0}[\delta_1(\hat{f}, \rho_n W_0)] \geq C \min \left( \sqrt{\frac{\rho_n}{n}}, \rho_n \right). \quad (99)$$

The proof of (97) follows from the proof of (43) in [26] using the trivial inequality

$$\|W_u(x, y) - W_v(\tau(x), \tau(y))\|_2^2 \leq \|W_u(x, y) - W_v(\tau(x), \tau(y))\|_1. \quad (100)$$

The proof of (98) follows the lines of the proof of (44) using that  $\|\mathbf{B}\|_2^2 = \|\mathbf{B}\|_1$  for matrices with entries in  $\{-1, 1\}$ . The proof of (99) is identical to the proof of (45) in [26].

In order to prove the upper bound (24), the proof of Proposition 3.2 in [26] can be easily modified to get an upper bound on the agnostic error measured in  $l_1$ -distance:

**Lemma 16** (Agnostic error measured in  $l_1$ -distance). *Consider the  $W$ -random graph model. For all integer  $k \leq n$ ,  $W_0 \in \mathcal{W}^+[k]$  and  $\rho_n > 0$ , we have*

$$\mathbb{E} [\delta_1(\tilde{f}_{\Theta_0}, f_0)] \leq C \rho_n \sqrt{\frac{k}{n}}.$$

Now (24) follows from Lemma 16 and (16). Finally, the  $\rho_n$  convergence rate is simply achieved by the constant estimator  $\hat{f} \equiv 0$ .

## H Proof of Proposition 7

For  $\Theta_0$  generated according to the sparse  $W$ -random graph model (25) with graphon  $W_0 \in \mathcal{W}_1^+$ , integrating (9) with respect to  $\xi$  and using  $\|W_0\|_1 = 1$ , we get

$$\mathbb{E}_{W_0} [\|A - \Theta_0\|_{\square}] \leq 6 \sqrt{\frac{\rho_n}{n}}.$$

So, using the triangle inequality (19) it is enough to bound the agnostic error  $\mathbb{E}_{W_0} [\delta_{\square}(\tilde{f}_{\Theta_0}, f'_0)]$ . We take  $W^* \in \mathcal{W}_1^+[k, \mu]$  (or  $W^* \in \mathcal{W}_2^+[k]$  in the case of  $L_2$  graphons) such that

$$\delta_1(W^*, W'_0) \leq \inf_{W \in \mathcal{W}_1^+[k, \mu]} \delta_1(W, W'_0) (1 + 1/n^2), \quad (101)$$

or  $\delta_2(W^*, W'_0) \leq \inf_{W \in \mathcal{W}_2^+[k]} \delta_2(W, W'_0) (1 + 1/n^2)$  for  $L_2$  graphons. Without loss of generality we can assume that  $\rho_n W^*(x, y) \leq 1$ . Let  $f^* = \rho_n W^*$  and  $\Theta^* = (\Theta_{ij}^*)$  be such such that for  $i \neq j$   $\Theta_{ij}^* = W^*[\xi_i, \xi_j]$  where  $(\xi_i)$  are the same as for  $\Theta_0$ . Triangle inequality implies

$$\begin{aligned} \mathbb{E}_{W_0} [\delta_{\square}(\tilde{f}_{\Theta_0}, f'_0)] &\leq \delta_{\square}(f'_0, f^*) + \mathbb{E}_{W_0} [\delta_{\square}(f^*, \tilde{f}_{\Theta^*})] + \mathbb{E}_{W_0} [\delta_{\square}(\tilde{f}_{\Theta^*}, \tilde{f}_{\Theta_0})] \\ &\leq 2\delta_1(f'_0, f^*) + \mathbb{E}_{W^*} [\delta_{\square}(f^*, \tilde{f}_{\Theta^*})] \end{aligned}$$

where we use  $\delta_{\square}(f'_0, f^*) \leq \delta_1(f'_0, f^*)$  and  $\mathbb{E}_{W_0} [\delta_{\square}(\tilde{f}_{\Theta^*}, \tilde{f}_{\Theta_0})] \leq \delta_1(f'_0, f^*)$  and that  $\tilde{f}_{\Theta^*}$  is distributed as under  $W^*$ . Similarly for  $L_2$  graphons, we obtain  $\mathbb{E}_{W_0} [\delta_{\square}(\tilde{f}_{\Theta_0}, f_0)] \leq 2\delta_2(f'_0, f^*) + \mathbb{E}_{W^*} [\delta_{\square}(f^*, \tilde{f}_{\Theta^*})]$ . Then, we use the following lemma:

**Lemma 17.** (i) *Consider any  $W^* \in \mathcal{W}_1^+[k, \mu]$  and  $\rho_n \geq 1/n$  such that  $\rho_n W^*(x, y) \leq 1$ . Then*

$$\mathbb{E}_{W^*} [\delta_{\square}(\tilde{f}_{\Theta^*}, f^*)] \leq C \left[ \rho_n \|W^*\|_1 \sqrt{\frac{k}{\mu n}} + \sqrt{\frac{\rho_n}{n}} \right].$$

(ii) Consider any  $W^* \in \mathcal{W}_2^+[k]$  and  $\rho_n \geq 1/n$  such that  $\rho_n W^*(x, y) \leq 1$ . Then,

$$\mathbb{E}_{W^*} \left[ \delta_{\square} \left( \tilde{f}_{\Theta^*}, f^* \right) \right] \leq C \left[ \rho_n \|W^*\|_2 \sqrt{\frac{k}{n}} + \sqrt{\frac{\rho_n}{n}} \right]. \quad (102)$$

Now (26) follows from (i) of Lemma 17 and  $\|W^*\|_1 \leq \|W_0\|_1(2+n^{-2})$ . The proof of (28) follows the same lines using (ii) of Lemma 17.

To prove (27) and (29) we only need to prove that  $\mathbb{E}_{W_0} [\|\tilde{\Theta}_\lambda - \Theta_0\|_{\square}] \leq C\sqrt{\rho_n/n}$ . Using the definition of  $\tilde{\Theta}_\lambda$  (13) we compute

$$\begin{aligned} \mathbb{E}_{W_0} [\|\tilde{\Theta}_\lambda - \Theta_0\|_{\square}] &\leq \mathbb{E}_{W_0} [\|A - \Theta_0\|_{\square}] + \mathbb{E}_{W_0} [\|\tilde{\Theta}_\lambda - A\|_{\square}] \\ &\leq 6\sqrt{\frac{\rho_n}{n}} + \mathbb{E}_{W_0} \left[ \frac{\|\tilde{\Theta}_\lambda - A\|_{2 \rightarrow 2}}{n} \right], \\ &\leq C\sqrt{\frac{\rho_n}{n}} \end{aligned}$$

where we used that  $\|B\|_{\square} \leq \|B\|_{2 \rightarrow 2}/n$  and the definition of  $\tilde{\Theta}_\lambda$ . This completes the proof of Proposition 7.

**Proof of Lemma 17.** Consider the matrix  $\Theta'$  with entries  $(\Theta')_{ij} = \rho_n W^*(\xi_i, \xi_j)$  for all  $i, j$ . As opposed to  $\Theta^*$ , the diagonal entries of  $\Theta'$  are not constrained to be null. By the triangle inequality, we get

$$\mathbb{E}_{W^*} [\delta_{\square} (\tilde{f}_{\Theta^*}, f^*)] \leq \mathbb{E}_{W^*} [\delta_{\square} (\tilde{f}_{\Theta^*}, \tilde{f}_{\Theta'})] + \mathbb{E}_{W^*} [\delta_{\square} (\tilde{f}_{\Theta'}, f^*)]. \quad (103)$$

Since the entries of  $\Theta^*$  coincide with those of  $\Theta'$  outside the diagonal, the difference  $\tilde{f}_{\Theta^*} - \tilde{f}_{\Theta'}$  is null outside of a set of measure  $1/n$ . Also, the entries of  $\Theta'$  are smaller than 1. It follows that  $\mathbb{E}[\delta_{\square}(\tilde{f}_{\Theta^*}, \tilde{f}_{\Theta'})] \leq 1/n \leq \sqrt{\rho_n/n}$ . Since  $\delta_{\square}(\tilde{f}_{\Theta'}, f^*) \leq \delta_1(\tilde{f}_{\Theta'}, f^*)$ , it suffices to prove that

$$\begin{aligned} \mathbb{E}_{W^*} [\delta_1(\tilde{f}_{\Theta'}, f^*)] &\leq C\rho_n \|W^*\|_1 \sqrt{\frac{k}{\mu n}}, \quad \text{for } W^* \in \mathcal{W}_1^+[k, \mu] \\ \mathbb{E}_{W^*} [\delta_1(\tilde{f}_{\Theta'}, f^*)] &\leq C\rho_n \|W^*\|_2 \sqrt{\frac{k}{n}}, \quad \text{for } W^* \in \mathcal{W}_2^+[k]. \end{aligned}$$

Since  $W^*$  is a  $k$ -step function, we can reorganize  $f^*$  and  $\tilde{f}_{\Theta'}$  in such a way that these two graphon are equal on a set of large Lebesgue value. More precisely, we adopt the same approach as in the proof of Theorem 1 and we only sketch the result here. Let  $\mathbf{Q} \in (\mathbb{R}^+)^{k \times k}_{sym}$  and  $\phi : [0, 1] \times [k]$  that characterize  $W^*$ . For  $a = 1, \dots, k$ , denote  $\lambda_a = \lambda(\phi^{-1}(\{a\}))$ . For any  $b \in [k]$ , define the cumulative distribution function  $F_\phi(b) = \sum_{a=1}^b \lambda_a$  and set  $F_\phi(0) = 0$ . For any  $(a, b) \in [k] \times [k]$  define  $\Pi_{ab}(\phi) = [F_\phi(a-1), F_\phi(a)] \times [F_\phi(b-1), F_\phi(b)]$ . Define  $W'(x, y) = \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \mathbb{1}_{\Pi_{ab}(\phi)}(x, y)$ . Obviously,  $f' = \rho_n W'$  is weakly isomorphic to  $f^* = \rho_n W^*$ . Now, let  $\hat{\lambda}_a = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\xi_i \in \phi^{-1}(a)\}}$  be the (unobserved) empirical frequency of group  $a$ . Consider a function  $\psi : [0, 1] \rightarrow [k]$  such that:

- (i)  $\psi(x) = a$  for all  $a \in [k]$  and  $x \in [F_\phi(a-1), F_\phi(a-1) + \hat{\lambda}_a \wedge \lambda_a)$ ,
- (ii)  $\lambda(\psi^{-1}(a)) = \hat{\lambda}_a$  for all  $a \in [k]$ .

Such a function  $\psi$  exists (for details see the Step 2 of the proof of Theorem 1). Finally define the graphon  $\hat{f}'(x, y) = \mathbf{Q}_{\psi(x), \psi(y)}$ . Notice that  $\hat{f}'$  is weakly isomorphic to the empirical graphon  $\tilde{f}_{\Theta^*}$ . Since  $\delta_1(\cdot, \cdot)$  is a metric on the quotient space of graphons, we have

$$\delta_1(\tilde{f}_{\Theta^*}, f^*) = \delta_1(\hat{f}', f') \leq \|\hat{f}' - f'\|_1.$$

The two functions  $f'(x, y)$  and  $\hat{f}'(x, y)$  are equal except possibly the case when either  $x$  or  $y$  belongs to one of the intervals  $[F_\phi(a-1) + \hat{\lambda}_a \wedge \lambda_a, F_\phi(a-1) + \lambda_a]$  for  $a \in [k]$  and we have

$$\begin{aligned} \|\hat{f}' - f'\|_1 &= \rho_n \left\| \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \mathbf{1}_{\Pi_{ab}(\phi)}(x, y) - \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \mathbf{1}_{\Pi_{ab}(\psi)}(x, y) \right\|_1 \\ &\leq \rho_n \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \lambda(\Pi_{ab}(\phi) \Delta \Pi_{ab}(\psi)) \\ &\leq \rho_n \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \left\{ |\lambda_a - \hat{\lambda}_a| \lambda_b + |\lambda_b - \hat{\lambda}_b| \lambda_a + |\lambda_a - \hat{\lambda}_a| |\lambda_b - \hat{\lambda}_b| \right\}. \end{aligned}$$

Since  $\xi_1, \dots, \xi_n$  are i.i.d. uniformly distributed random variables,  $n\hat{\lambda}_a$  has a binomial distribution with parameters  $(n, \lambda_a)$ . By Cauchy-Schwarz inequality we get  $\mathbb{E}[|\lambda_a - \hat{\lambda}_a|] \leq \sqrt{\lambda_a(1 - \lambda_a)/n}$  and  $\mathbb{E}[|\lambda_a - \hat{\lambda}_a| |\lambda_b - \hat{\lambda}_b|] \leq \sqrt{\lambda_a \lambda_b / n}$ . Then, we get

$$\mathbb{E}_{W^*} \|\hat{f}' - f'\|_1 \leq \frac{\rho_n}{\sqrt{n}} \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \left\{ \sqrt{\lambda_a} \lambda_b + \sqrt{\lambda_b} \lambda_a + \sqrt{\frac{\lambda_a \lambda_b}{n}} \right\}.$$

Now for  $W^* \in \mathcal{W}_1^+[k, \mu]$  we use  $\lambda_a \geq \mu/k$  for all  $a \in [k]$  to get

$$\mathbb{E}_{W^*} \|\hat{f}' - f'\|_1 \leq C \rho_n \sqrt{\frac{k}{\mu n}} \sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab} \lambda_a \lambda_b (1 + \sqrt{\frac{k}{\mu n}}) = C \rho_n \|W^*\|_1 \sqrt{\frac{k}{n}},$$

since we assume that  $k \leq \mu n$ . For  $W^* \in \mathcal{W}_2^+[k]$  we use the Cauchy-Schwarz inequality:

$$\mathbb{E}_{W^*} \|\hat{f}' - f'\|_1 \leq \frac{2\rho_n}{\sqrt{n}} \sqrt{\sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab}^2 \lambda_a \lambda_b} \sqrt{\sum_{a=1}^k \sum_{b=1}^k \lambda_b} + \frac{\rho_n k}{n} \sqrt{\sum_{a=1}^k \sum_{b=1}^k \mathbf{Q}_{ab}^2 \lambda_a \lambda_b} = C \rho_n \|W^*\|_2 \sqrt{\frac{k}{n}},$$

since  $k \leq n$ . □

## References

- [1] Edo M Airolidi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.
- [2] Noga Alon, W. Fernandez De La Vega, Ravi Kannan, and Marek Karpinski. Random sampling and approximation of max-csps. *Journal of computer and system sciences*, 67(2):212–243, 2003.
- [3] Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 2016.

- [4] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.
- [5] Peter J Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.
- [6] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, 31(1):3–122, 2007.
- [7] C. Borgs, J.T. Chayes, H. Cohn, and S. Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *ArXiv e-prints*, August 2015.
- [8] C. Borgs, J.T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.*, 219(6):1801–1851, 2008.
- [9] C. Borgs, J.T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Ann. of Math. (2)*, 176(1):151–219, 2012.
- [10] Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pages 1369–1377, 2015.
- [11] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An lp theory of sparse graph convergence i: limits, sparse random graph models, and power law distributions. *arXiv preprint arXiv:1401.2906*, 2014.
- [12] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An lp theory of sparse graph convergence ii: Ld convergence, quotients, and right convergence. *arXiv preprint arXiv:1408.0744*, 2014.
- [13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003.
- [14] Diana Cai, Nathanael Ackerman, and Cameron Freer. An iterative step-function estimator for graphons. *arXiv preprint arXiv:1412.2129*, 2014.
- [15] Stanley H. Chan and Edoardo M. Airoldi. A consistent histogram estimator for exchangeable graph models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 208–216, 2014.
- [16] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015.
- [17] D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.
- [18] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, 2007.
- [19] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.
- [20] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.
- [21] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck’s inequality. *Probability Theory and Related Fields*, 165(3):1025–1049, 2016.
- [22] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.
- [23] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.
- [24] Svante Janson. *Graphons, cut norm and distance, couplings and rearrangements*, volume 4 of *New York Journal of Mathematics. NYJM Monographs*. State University of New York,

- University at Albany, Albany, NY, 2013.
- [25] Olga Klopp. Rank penalized estimators for high-dimensional matrices. *Electron. J. Statist.*, 5:1161–1183, 2011.
  - [26] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.*, 45(1):316–354, 2017.
  - [27] Pierre Latouche and Stéphane Robin. Bayesian model averaging of stochastic block models to estimate the graphon function and motif frequencies in a w-graph model. Technical report, Technical report, 2013.
  - [28] László Lovász. *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012.
  - [29] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *J. Combin. Theory Ser. B*, 96(6):933–957, 2006.
  - [30] S. J. Szarek. On the best constants in the Khinchin inequality. *Studia Math.*, 58(2):197–208, 1976.
  - [31] Endre Szemerédi. Regular partitions of graphs. Technical report, DTIC Document, 1975.
  - [32] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.
  - [33] Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.
  - [34] Justin Yang, Christina Han, and Edoardo M. Airoldi. Nonparametric estimation and testing of exchangeable graph models. In *AISTATS*, 2014.
  - [35] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.